

# A Theory of Policy Sabotage

Alexander V. Hirsch\*  
Division of the Humanities  
and Social Sciences  
California Institute of Technology  
avhirsch@hss.caltech.edu

Jonathan P. Kastellec  
Department of Politics  
Princeton University  
jkastell@princeton.edu

December 22, 2021

## Abstract

We develop a theory of policymaking that examines when policy sabotage—the deliberate choice by an opposition party to interfere with the implementation of a policy—can be an effective electoral strategy, even if rational voters can observe that it is happening. In our model, a potential saboteur chooses whether to sabotage an incumbent’s policy by blocking its successful implementation. A voter then decides whether to retain the incumbent, who is of unknown ability, or to select a challenger. We find that the incentives for sabotage are broadly shaped by the underlying popularity of the incumbent—it is most attractive when an incumbent is somewhat unpopular. If so, sabotage may decrease the probability the incumbent is reelected, even though sabotage is observable to the voter. This is because while the saboteur knows that sabotage will improve the incumbent’s reputation, he fears that absent sabotage a policy success will improve that reputation even more.

---

\*We thank Peter Buisseret, Brandice Canes-Wrone, Laura Doval, Saba Devdariani, Sandy Gordon, Ian Turner, and Craig Volden, and seminar audiences at Emory, New York University Law School, and Yale for helpful comments and suggestions.

A central tension in democratic theory concerns how imperfectly informed voters can either select representatives who act in their best interest, or sanction representatives who do not (Fearon 1999). To address this tension, scholars have developed an extensive literature that employs the theory of political agency to understand how and why reelection-minded representatives may choose to act in the best interests of voters, even if voters can only imperfectly observe whether representatives are actually doing so.<sup>1</sup>

While the nuances of various theories differ, a ubiquitous theme of models of selection and accountability is that voters condition their retention choices on the observable actions of politicians. This makes perfect sense, as voters should use all available information at their disposal—in particular, policy outcomes. Given this, in a world in which one party seamlessly controls policy (such as in a parliamentary system with a sizable majority party), we would expect that party to avoid observable actions (as opposed to hidden ones) that decrease the chance of a successful policy outcome.

However, in a context where power is more fragmented, either because of institutional-based gridlock and/or party-based polarization, the motivations of competing parties are more complicated. In particular, the current era of partisan polarization in the United States has seen an apparent increase in the incidence of politicians engaging in “policy sabotage”—the deliberate effort to hinder the implementation of a policy enacted by the opposition party. For example, since 2010 congressional Republicans have sought to undermine the implementation of the Affordable Care Act (“Obamacare”)—an effort joined with full gusto by President Trump when he took office in 2017—and have not been shy about their intentions.<sup>2</sup> More generally, Lee (2016) argues that the increase in competitiveness for control of Congress has disincentivized minority parties from working on policy in a bipartisan fashion, and has instead encouraged a focus on activities that hurt the reputation of

---

<sup>1</sup>See Ashworth (2012) for an outstanding review of this literature.

<sup>2</sup> In 2017, for example, the Trump administration cut off billions of dollars in subsidies to insurers who enroll Americans through Obamacare—the subsidies were designed to help lower out-of-pocket expenses for low-income enrollees in the program (Pear, Haberman and Abelson 2017).

the party in power, such as “messaging” legislation.<sup>3</sup>

From the perspective of theoretical models of accountability, policy sabotage poses a puzzle: why is sabotage a (potentially) effective strategy for damaging a party’s electoral prospects when voters can *see* it and update on its deployment as a strategy? In this paper we present a formal theory of policy sabotage that examines this question. We develop a two-period model in which a voter chooses to either reelect an incumbent or replace him with a challenger. Incumbents and challengers are each associated with an ideology, and can be either low or high ability. This ability, which is not known to the voters, affects the probability that a policy they generate will translate into a successful outcome.

The key innovation of the model is that there is a potential “saboteur” who can interfere with implementation of the current officeholder’s policy. The saboteur can be conceptualized as a bureaucrat or an out-party, depending on the context. Specifically, the saboteur can choose to let the policy be implemented, which means that it will succeed with some positive probability that is based on the incumbent’s ability. Alternatively, the saboteur can sabotage the policy, which will ensure failure. Importantly, unlike standard agency models with effort (e.g. Ashworth and Bueno de Mesquita (2017)), we assume that both implementing and sabotaging the policy are costless; this means the decision of whether to engage in sabotage is *not* one related to costly effort. In addition, the act of sabotage is perfectly observable to the voter, as was arguably the case for the Affordable Care Act.<sup>4</sup>

The voter and the saboteur are the strategic players in the model. For simplicity, we assume that the incumbent and challenger are non-strategic, and passively committed to

---

<sup>3</sup>We return to the connection between Lee’s work and our model in the discussion section below.

<sup>4</sup>A potential objection to our approach is that sabotage is often unobservable, particularly when conducted by career bureaucrats deep in the executive branch. While we agree that unobservable sabotage is a question of interest, we also argue that modeling it directly is unlikely to lead to many (if any) interesting results given how straightforward the saboteur’s incentives would be. In footnote 13 below, we discuss in greater detail how our results would change if sabotage were unobservable. By assuming sabotage is observable, we set up a more difficult test for it to be effective as a political strategy.

generating the best possible policies that align with their personal ideologies. Thus, the likelihood of a policy success depends on the officeholder’s intrinsic ability, rather than any sort of strategic decision. (In practice, this simplification means the model is one of selection with respect to the incumbent policymaker, not accountability.<sup>5</sup>)

Importantly, in addition to ideology, both the voter and the saboteur have a shared preference for successful outcomes. The voter is assumed to be imperfectly informed about the incumbent’s ability, but can learn more by observing the outcome of his policy in each period (success or failure) as well as the saboteur’s decision over whether to engage in sabotage. Following the realization of the outcome in the first period, the voter chooses whether to retain the incumbent or replace him with a challenger. The game then repeats in the second period.

We present two versions of the model that differ in the information available to the saboteur. In the first version, the saboteur does not know the incumbent’s ability, and thus holds the same uncertainty as the voter. We call this an *uninformed saboteur*. In the second version, the saboteur knows the incumbent’s ability. This assumption, which we denote the *informed saboteur*, is more realistic if one believes that actors in government are better informed than voters about the ability of fellow policymakers.

The saboteur’s dilemma is as follows. Assume he is ideologically closer to the challenger than the incumbent. Because sabotage ensures policy failure, the saboteur may choose to sabotage because it prevents the voter from learning about the incumbent’s ability from policy outcomes. Such a blocking maneuver *may* increase the chance that the voter chooses to replace the incumbent with the challenger—but only under circumstances we discuss

---

<sup>5</sup> Our assumptions that both the incumbent and challenger are non-strategic rules out some potential interesting strategic possibilities, such as the incumbent moderating his policy positions to signal ability. We acknowledge this limited scope of the model. By placing the strategic decisions directly in the hands of the saboteur, our goal is to highlight the strategic incentives facing a potential saboteur, and thus for simplicity we model a saboteur who does not face an election, but whose actions potentially affect the electoral fates of the incumbent and the challenger. We believe our model provides a foundation for future work to more directly incorporate the strategic choices of incumbents and challengers.

shortly. However, the fact that the saboteur also cares about the success of the policy may push him in the opposite direction—especially because implementing the policy is costless.

The voter’s dilemma is as follows. Suppose first that the voter believes the saboteur to be uninformed about the incumbent’s ability, which is the simpler case. Then observing a policy success (which, recall, can only occur in the absence of sabotage) increases the voter’s belief that the incumbent is of high ability—but only probabilistically. Sabotage, on the other hand, prevents any learning from occurring because it blocks successful implementation with certainty. Sabotage will thus induce the voter to replace an incumbent she might have otherwise retained if and only if the incumbent is *somewhat unpopular*—that is, if the voter is inclined to replace the incumbent initially, but would have been willing to retain him after observing a policy success. Whether the saboteur actually chooses to sabotage such an incumbent, in turn, depends on whether he is willing to sacrifice a policy success to *suppress* the voter’s ability to learn more about the incumbent’s ability through policy outcomes.

The voter’s inferences, and thus the resulting equilibrium, are more complex when she believes the saboteur to be privately informed about the incumbent’s ability. In this case, sabotage (and its absence) can *itself* signal information about the incumbent’s ability. If, for example, the voter believes the saboteur to be sabotaging a high-ability incumbent to block the voter from learning about the incumbent’s ability, then she will infer from *sabotage itself* that the incumbent is high ability and reelect him; thus, sabotage will backfire. Conversely, if the voter believes the saboteur to be sabotaging a low-ability incumbent to signal that they are low ability, then sabotage will harm the incumbent’s prospects, which will incentivize the saboteur to sabotage the incumbent regardless of their ability. Thus, it is not obvious *a priori* what a rational voter will infer about the incumbent’s ability when sabotage occurs. Indeed, our analysis uncovers two particularly interesting equilibria that illustrate how sabotage can both communicate different information, and have different electoral effects.

The first equilibrium prevails when the incumbent is somewhat unpopular. In it, the saboteur sometimes sabotages both a high- and low-ability incumbent, but surprisingly, is

more likely to sabotage a high-ability one. Sabotage thus credibly signals to the voters that the incumbent is high ability, which perversely *helps* his reputation and his electoral prospects. Conversely, the absence of sabotage *hurts* the incumbent’s reputation—but not so much that a policy success cannot overcome that harm and carry him to reelection with a high likelihood. How can it be that a rational saboteur undertakes sabotage even though it helps the incumbent’s reputation? It is not because the saboteur thinks sabotage will harm the incumbent’s reputation—he anticipates that rational voters will see through such a strategy. Rather, it is because the saboteur knows that sabotage improves the incumbent’s reputation, but fears that absent sabotage a policy success will improve that reputation even more.

The second equilibrium prevails when the incumbent is *very popular*—that is, when a policy failure would not induce replacement absent additional information. In this equilibrium, the saboteur sometimes sabotages a high-ability incumbent, but *always* sabotages a low-ability one. Sabotage thus credibly, but imperfectly, signals to the voter that the incumbent is low ability, which harms his reputation and electoral prospects. Paradoxically, the saboteur’s ability to credibly harm the incumbent’s electoral prospects via sabotage is precisely due to the incumbent’s initial popularity. When the incumbent is starting out so far ahead that the voter will retain him even after failure, the saboteur has no greater *electoral* incentive to sabotage a high-ability incumbent than a low-ability one—the former is more likely to generate a policy success, but both will still be retained even if they fail. However, there remains a greater intrinsic *cost* to sabotaging a high-ability incumbent, because it is more likely to prevent a policy success from which even the saboteur would benefit. This greater cost allows the saboteur to credibly signal the incumbent’s low ability via sabotage.

**Related literature** Our paper is connected to several literatures. First, the act of sabotage can be seen as somewhat akin to a veto in canonical models of “veto politics” in presidential systems, particularly the United States (Cameron 2000, Groseclose and McCarty 2001). As with a presidential veto, the saboteur in our model unilaterally halts the

policy-making process. However, such models typically analyze the effect of uncertainty about the *preferences* of the relevant actors, rather than their *abilities*, as in our model. In this respect, our theory more closely resembles Buisseret (2016). In both his and our model, the proposer is of variable ability, which is (sometimes) known to the veto player. This ability, in turn, probabilistically determines whether a policy succeeds or not. Our focus, however, is different from Buisseret, who examines the differences between political systems in which competing factions are jointly appointed by voters, such as parliamentary systems, or political systems in which competing factions are separately appointed, such as in presidential systems. He finds that joint appointment institutions reduce the incentives for the veto player to engage in obstruction, thereby improving voter welfare. Conversely, in separate systems—as in the United States—veto players are more incentivized to engage in obstruction in order to establish a reputation for competence. Despite the technical and substantive differences between the two, our model can be seen as extending hBuisseret’s insights to address when policy obstruction is rational for a competing party or politician.

One important conceptual difference, however, between presidential vetoes in the U.S. system and sabotage is that whereas the former embodies “institutionalized” policy blocking, sabotage generally captures more informal and diffuse mechanisms of blockage, such as limited or half-hearted policy implementation. In addition, it is not clear whether it makes sense to conceptualize a formalized veto itself as “sabotage,” given the president’s constitutionally delineated role in bill approval.

Along these lines, our theory is also related to recent work on what opposition parties gain from engaging in the tactics of delay and obstruction (Patty 2016, Fong and Krehbiel 2018, Gieczewski and Li 2021). However, in these models the opposition does not actually affect the ultimate implementation of policy, only its timing. Gieczewski and Li (2021) complements our work by studying the optimal timing of obstruction over the course of an incumbent’s term, and shares our key insight that it is incumbents whose standing is *moderate* with the voters that invite sabotage. However, our model has several key differences. First,

our politicians are policy rather than office motivated, and share a preference for successful outcomes that they balance against the ideological benefits of electoral turnover. Second, we model sabotage as costless in terms of effort but genuinely destructive in terms of policy consequences, rather than costly in terms of effort and merely obstructive in terms of policy consequences. These differences are crucial for the equilibrium predictions of our main model, in which the saboteur is better-informed about the incumbent’s ability than the voter, and so the act of sabotage carries “signaling” implications that depend on the saboteur’s policy preferences.

Our paper is also connected to several related models in the broad literature on democratic accountability (e.g. Ashworth and Bueno De Mesquita 2014). In particular, our model complements research on the role of challengers in democratic accountability. As Shotts and Ashworth (2011) note, most theories of accountability feature “passive” challengers who exist as alternatives to the incumbent (see e.g. Ferejohn 1986, Gordon and Huber 2002, Maskin and Tirole 2004, Ashworth and Bueno De Mesquita 2008; 2014). In other papers, challengers do take affirmative yet limited actions, such as entering the race as an alternative to the incumbent and/or declaring a competing platform (see e.g. Epstein and Zemsky 1995, Gordon, Huber and Landa 2007). In contrast, Shotts and Ashworth (2011) develop a model in which challengers can make statements to voters about which of two policies is “correct” (in the sense of matching the true state of the world)—in some equilibria, these statements can affect whether the incumbent is retained (see also Lemon 2005, Warren 2012). Our model follows most of the literature in assuming a passive challenger. However, in some instances the saboteur works “on behalf of” the challenger in the hopes of defeating the incumbent—our theory can thus be placed in a broader class of models where actors take affirmative steps to try to bolster the chance of challenger victory.

In addition, there is a connection between our results and a phenomenon called in the political agency literature called “gambling for resurrection” (Downs and Rocke 1994). This occurs when a weak incumbent—that is, one who is somewhat unpopular in the language



of our model—takes a risky action in the hopes that it will turn out well and get them over the electoral threshold (Dewan and Hortala-Vallve 2017, Izzo 2020). In some sense, sabotage with an uninformed saboteur is the flipside of this—although the policy is exogenous, the equilibrium can be interpreted as an intermediary trying to intervene to *prevent* the incumbent from “gambling” that a policy success will carry him to reelection.<sup>6</sup>

In terms of its assumptions about the preferences of politicians and voters, our theory connects to a burgeoning literature analyzing models in which policies have a valence or quality component that all players value despite their ideological differences (Lax and Cameron 2007, Ting 2009, Hirsch and Shotts 2012; 2015, Hitt, Volden and Wiseman 2017, Turner 2017, Hirsch and Shotts 2018). This assumption often invites skepticism on the grounds that politicians’ intrinsic desire for the failure of ideologically distant-policies seems self-evident. Our model explores the plausible alternative that this desire is rooted in the *strategic* incentive to expedite ideological policy change, which in turn requires an understanding of the conditions under which open sabotage can facilitate such change. Returning to the Affordable Care Act, a real-world example of our underlying assumptions is that pivotal Republican decision-makers (who strongly opposed the passage of the law) still somewhat valued “good policy” considerations, such as increases in coverage and reductions in cost—and thus would have preferred the ACA to achieve these outcomes rather than not *if there was no chance that it would ever be repealed*. Similar to our assumption about sabotage being observable, we believe that this “valence” assumption establishes an interesting strategic dynamic through which to explore the sabotage decision, as opposed to simply assuming that politicians intrinsically want opposition policies to fail.

Finally, our model speaks to the burgeoning empirical literature on “blame attribution,” which evaluates how citizens appropriate blame across policy-makers in the wake of policy

---

<sup>6</sup>Another related finding occurs in Canes-Wrone, Herron and Shotts (2001), in which moderately popular incumbents will pander to public opinion only if it is unlikely the public discovers that the incumbent pandered before the election; in other words, the incumbent “gambles” that pandering will not be detected in time to affect the election.

failures (Healy and Malhotra 2013, 291-3). Much of this research focuses on how partisan cues may bias citizen evaluation of the actions of elected officials, particularly when blame may be plausibly distributed across multiple parties (as occurs frequently in a system of federalism) (see e.g. Arceneaux 2005; 2006, Malhotra and Kuo 2008, Healy, Kuo and Malhotra 2014). While there are no parties as such in our model, as we noted one interpretation of the saboteur is that of the out-party who can block implementation. Our model illustrates that the effects of sabotage, as well as the incentives to engage therein, are rich and multifaceted even when voters have no difficulty attributing blame for policy failures.

## 2 The Model

We model a game played by two policy-motivated actors; a decisionmaker  $V$ , interpreted as a voter, and a potential saboteur  $S$  (henceforth just saboteur), interpreted as an unelected actor in government who can influence policy-implementation by elected officeholders. At the start of the game, there is an incumbent politician and a challenger politician, denoted  $j \in J = \{I, C\}$  respectively. (For presentational clarity, we use male pronouns to refer to the saboteur and female pronouns to refer to the voter; at times we refer to the incumbent or challenger using plural pronouns.) Each politician is associated with an ideal spatial ideology  $x_j \in [-\infty, \infty]$ ; we assume without loss of generality that  $x_I > x_C$ . In addition, however, the incumbent and the challenger may be either of “low” or “high” ability, as denoted by  $\lambda_j \in \{L, H\}$ . This type affects the likelihood that a policy they generate will result in a successful outcome. For simplicity, we abstract away from strategic policy decisions by politicians. The incumbent and challenger, who are exogenous and are not players in the game, represent elected politicians that passively generate policies whose ideologies matches their own, and which succeed with a likelihood associated with their ability.

The voter and the saboteur each have an ideological ideal point  $x_i$  and suffer spatial loss from the distance of the current policy’s ideology to their own ideal point. However, the players also have a shared preference for successful outcomes. Specifically, a policy generated by officeholder  $j^t$  in period  $t$  must be implemented with “effort”  $e^t \in \{0, 1\}$ . Although we

start with the terminology of effort to clarify the connection with standard principal-agent models, a key assumption in our model is that effort is actually free. Accordingly, “effort” should be interpreted as letting the officeholder’s policy run its natural course, while “no effort” means to actively interfere with its success via sabotage. We also use “implementing” and “sabotaging” to refer to effort and its absence, respectively.

Policymaking in each period  $t \in \{1, 2\}$  proceeds as follows. First, the saboteur observes a private signal  $\eta^t \in \{L, H, \emptyset\}$  that is equal to the current officeholder’s true ability  $\lambda_{j^t}$  with probability  $\rho$  and equal to  $\emptyset$  (i.e., is uninformative) with probability  $1 - \rho$ .<sup>7</sup> Following this, the current officeholder  $j^t$  generates a policy. Next, the saboteur chooses an effort level  $e^t \in \{0, 1\}$ . Finally, the officeholder’s policy either succeeds or fails ( $y^t \in \{0, 1\}$ ). Player  $i$  places a value  $\gamma_i > 0$  on success, so both players value successes. A policy generated by officeholder  $j^t$  in period  $t$  that is implemented with effort  $e^t$  succeeds with probability  $e^t \cdot q_{\lambda_{j^t}}$ . Thus,  $0 < q_L < q_H < 1$  represent the probabilities that policies generated by each type of politician will succeed if implemented, while sabotage ensures failure. Finally, politician  $j \in \{I, C\}$  is high ability with prior probability  $\theta_j \in [0, 1]$  (the abilities of the incumbent and challenger are uncorrelated, and each may have different prior probabilities of being high ability).

For our main results we consider two specific versions of the model in which the information available to the saboteur about the current officeholder’s ability differs. In the first, the saboteur is *always uninformed* about the officeholder’s ability ( $\rho = 0$ ). Thus, while his decision to sabotage will affect what success and failure reveal about the officeholder’s ability, it cannot itself signal information to the voter. In the second, the saboteur is *perfectly informed* about the current officeholder’s ability ( $\rho = 1$ ). Thus, his decision to sabotage—in

---

<sup>7</sup>For simplicity, we assume that the saboteur may only become informed about the incumbent’s ability. Allowing the saboteur to also become informed about the challenger’s ability would not undermine the model’s central tension – that sabotaging a high-ability incumbent is both intrinsically costlier and potentially more electorally effective. However, equilibrium strategies would be complicated by the fact that the saboteur will always be more willing to sabotage when the challenger is high ability *ceteris-paribus*, implying that sabotage (its absence) may signal additional positive (negative) information about the challenger.

Parameter	Description
$V$	Voter
$S$	(Potential) Saboteur
$j \in J = \{I, C\}$	Incumbent and Challenger
$x_j \in [-\infty, \infty]$	Spatial ideology of each politician's policies
$t \in \{1, 2\}$	Period
$\lambda_j \in \{L, H\}$	Low-ability or high-ability politician
$e^t \in \{0, 1\}$	"Effort" of saboteur (0 = sabotage, 1 = not sabotage)
$y^t \in \{0, 1\}$	Failure or Success of Policy
$\gamma_i > 0$	Value of success to player $i$ (Valence)
$e^t \cdot q_{\lambda_{j^t}}$	Probability of success ( $0 < q_L < q_H < 1$ )
$\theta_j \in [0, 1]$	Prior probability politician $j \in \{I, C\}$ is high ability
$U(x_i; x_I, x_C)$	Net policy benefit
$V(\hat{\theta}_I, \hat{\theta}_C; \gamma_i, q)$	Net valence benefit
$\Delta_{\lambda_i}(\pi_e^y)$	Impact probability for type $\lambda_I$

Table 1: Summary of notation

addition to influencing the probabilities of success and failure—may also directly signal his private information about that ability.

Finally, players' utility over the two periods is the discounted sum based on the ideology of the promulgated policies and their outcomes, i.e.

$$\sum_{t=1}^2 \delta^{t-1} \cdot (-(x_i - x_{j^t})^2 + \gamma_i \cdot y^t),$$

where  $x_{j^t}$  denotes the ideological location of the policy generated in period  $t$  and  $y^t$  denotes the outcome in period  $t$ . Table 1 summarizes the model's notation.

**Sequence of play** The game proceeds as follows.

1. Nature selects the incumbent's ability.
2. The saboteur observes a private signal  $\eta^t \in \{L, H, \emptyset\}$  of the incumbent's ability whose conditional distribution depends on the model variant.
3. The saboteur chooses whether to implement  $e^t \in \{0, 1\}$  the policy generated by officeholder  $j^t \in \{I, C\}$ ; this implementation choice (sabotage or not) is *observable* to the voter/decisionmaker.
4. The policy outcome ( $y^1 \in \{0, 1\}$ ) is realized; this is also observable to the voter.
5. The voter decides to retain the incumbent ( $j^2 = I$ ) or switch to the challenger ( $j^2 = C$ ).

6. The second round of play occurs, and steps (1)-(4) repeat.
7. The game ends when the second period policy is realized ( $y^2 \in \{0, 1\}$ ).

### 3 Preliminary Analysis

#### 3.1 Second Period

In the second period there is no impending election. The saboteur thus always implements the policy ( $e^2 = 1$ ) regardless of his beliefs about the incumbent's ability; this is because effort is free, an implemented policy will succeed with strictly positive probability, and the saboteur values success. Thus, from the perspective of an arbitrary player  $i$  with ideal point  $x_i$  and interim beliefs  $\hat{\theta}_j$  about politician  $j$ 's ability at the end of the first period (beliefs that are computed using Bayes' rule and the equilibrium strategies), the expected future payoff from having politician  $j \in \{I, C\}$  in office for the second period is

$$\gamma_i \left( q_L + \hat{\theta}_j (q_H - q_L) \right) - (x_i - x_j)^2.$$

Consequently, the second period *net benefit* of retaining the incumbent  $I$  rather switching to the challenger  $C$  consists of both a *net ideological benefit* and a *net valence benefit*:

$$\underbrace{(x_i - x_C)^2 - (x_i - x_I)^2}_{\text{net ideological benefit}} + \underbrace{\gamma_i (\hat{\theta}_I - \hat{\theta}_C) (q_H - q_L)}_{\text{net valence benefit}}$$

We denote the net ideological benefit as  $U(x_i; x_I, x_C)$ . It is increasing in the ideological alignment of player  $i$  with the incumbent, and is positive iff  $x_i > \frac{x_I + x_C}{2}$  (recall we have assumed  $x_I > x_C$ ). We denote the net valence benefit as  $V(\hat{\theta}_I, \hat{\theta}_C; \gamma_i, q)$  (where  $q = \{q_L, q_H\}$ )—this is increasing in player  $i$ 's value for success  $\gamma_i$ , and in the difference in her interim beliefs  $\hat{\theta}_I - \hat{\theta}_C$  about the abilities of the incumbent and the challenger. It is negative if the challenger is believed to be higher ability than the incumbent. It is also increasing in  $q_H - q_L$ , the difference in the probability that a high- vs. low-ability incumbent generates a policy that succeeds absent sabotage.

## 3.2 First Period

We now characterize first period play. The strategies of the two players take the following form:

- **Saboteur:** The saboteur's strategy is a probability of exerting effort  $e_\eta \in [0, 1]$  as a function of his private signal  $\eta \in \{L, H, \emptyset\}$  about the incumbent ability  $\lambda_I$ .
- **Voter:** The voter's strategy is a probability of retaining the incumbent  $\pi_e^y \in [0, 1]$  as a function of the saboteur's observed effort level  $e$  and the observed outcome  $y$ .

### 3.2.1 The Saboteur's Calculus

The saboteur's willingness to implement or sabotage depends on: (1) the effect of implementing on contemporaneous success, (2) the net future benefit of retaining the incumbent, and (3) the effect of implementing on the probability that the incumbent is retained.

**Effect of implementation on first-period success** After observing a signal  $\eta \in \{L, H, \emptyset\}$  about the incumbent's ability  $\lambda_I$ , the saboteur forms a posterior belief  $\Pr(\lambda_I = H|\eta) = \theta_I^\eta$  that the incumbent is high ability. He therefore assesses the likelihood of success after implementing the incumbent policy to be  $q_\eta$ , where

$$q_\eta = \theta_I^\eta \cdot q_H + (1 - \theta_I^\eta) \cdot q_L.$$

Hence, the net valence benefit of implementation is  $q_\eta \gamma_S$ .

**Net benefit of retaining incumbent** The saboteur's expected net benefit from the incumbent being reelected is  $U(x_S; x_I, x_C) + V(\theta_I^\eta, \theta_C; \gamma_S, q)$  since his posterior belief about the incumbent's ability is  $\theta_I^\eta$ .

**Effect of implementation on retention probabilities** The saboteur observes a private signal  $\eta \in \{L, H, \emptyset\}$  about the incumbent's ability  $\lambda_I$  and also has beliefs (that are correct in equilibrium) about the probability the voter will retain the incumbent  $\pi_e^y$  down each path of play. He can thus calculate how much implementing will affect the probability that the voter retains the incumbent, which crucially influences his willingness to sabotage.

Because the probability of success depends on the incumbent ability  $\lambda_I$ , so too does the impact of implementation on the probability of retention. We henceforth call this quantity the *impact probability* after signal  $\eta$ , and denote it as  $\Delta_\eta(\pi_e^y)$ .

Should the saboteur engage in sabotage ( $e = 0$ ), failure will result for sure and the incumbent will be retained with probability  $\pi_0$ . If he instead implements the incumbent's policy ( $e = 1$ ), he expects it to succeed and fail with probabilities  $q_\eta$  and  $1 - q_\eta$ , respectively; the incumbent will therefore be retained and replaced with probabilities  $\pi_1^1$  and  $\pi_1^0$ , respectively. The impact probability after receiving signal  $\eta$  is thus:

$$\Delta_\eta(\pi_e^y) = (q_\eta \cdot \pi_1^1 + (1 - q_\eta) \cdot \pi_1^0) - \pi_0^0 = (\pi_1^0 - \pi_0^0) + q_\eta(\pi_1^1 - \pi_1^0)$$

**Total net benefit** Combining the preceding observations, the net benefit to the saboteur of implementation after receiving signal  $\eta \in \{L, H, \emptyset\}$  is:

$$q_\eta \gamma_S + \delta \Delta_\eta(\cdot) (V(\theta_I^\eta, \theta_C; \gamma_S, q) + U(x_S; x_I, x_C)).$$

Implementing the incumbent's policy after signal  $\eta$  is a best response i.f.f. this quantity is  $\geq 0$ , and sabotage is a best response i.f.f. this quantity is  $\leq 0$ .

In our analysis, we focus on the specific case of a saboteur whose relative ideological preference for the challenger is sufficiently strong that he would prefer to sabotage even a relatively skilled incumbent if it would have sufficient electoral impact.

**Assumption 1** Assume  $q_H \gamma_S + \delta q_H (V(1, 0; \gamma_S, q_L = 0) + U(x_S; x_I, x_C)) < 0$

$$\iff \frac{-U(x_S; x_I, x_C)}{\gamma_S} > \frac{1}{\delta} + q_H$$

Formally, Assumption 1 states that the saboteur prefers to sabotage a known high-ability incumbent ( $\eta = H \rightarrow \theta_I^\eta = 1$ ) facing a known low-ability challenger ( $\theta_C = 0$ ) who is also certain to fail ( $q_L = 0$ ), as long as the resulting decrease  $\Delta_H$  in the incumbent's retention probability is at least  $q_H$ . When Assumption 1 holds, there is a unique strictly interior impact probability  $\bar{\Delta}_\eta(\cdot) \in (0, 1)$  for each signal  $\eta \in \{L, H, \emptyset\}$  of the incumbent's ability:

$$\bar{\Delta}_\eta(\cdot) = \frac{q_\eta}{\delta \left( \frac{-U(x_S; x_I, x_C)}{\gamma_S} - (q_H - q_L)(\theta_I^\eta - \theta_C) \right)} \quad (1)$$

above which the saboteur would strictly prefer to sabotage an incumbent of that type, and below which he would not. Because  $q_H > q_L$  (a high-ability incumbent is more likely to succeed “today”) and  $V(1, \theta_C; \gamma_S, q) > V(0, \theta_C; \gamma_S, q)$  (high-ability office-holders are more likely to succeed “tomorrow”), it is straightforward that  $\bar{\Delta}_H(\cdot) > \bar{\Delta}_\emptyset(\cdot) > \bar{\Delta}_L(\cdot)$ . That is, the stronger is the saboteur’s signal of the incumbent’s ability, the higher the electoral impact of sabotage must be to induce sabotage.

### 3.2.2 The Voter’s Calculus

When the voter makes her retention decision, she has already formed interim beliefs about the incumbent’s ability, which we denote  $\tilde{\theta}_I^{e,y}(e_L, e_H, e_\emptyset)$ . These beliefs are calculated from Bayes’ rule whenever possible, and are based on two observable actions: whether or not the saboteur exerted effort  $e$ , and the policy outcome  $y$  (success or failure). The beliefs also depend what the voter thinks about the saboteur’s unobserved strategy  $(e_L, e_H, e_\emptyset)$ —that is, the likelihood he that exerts effort after each signal  $\eta \in \{L, H, \emptyset\}$  of the incumbent’s ability. (The voter’s beliefs about the challenger policy remain at the prior  $\theta_C$  since the saboteur is known to be uninformed about the ability of policymakers out of office.) The voter decides whether to retain the incumbent based on these beliefs. We examine the retention decision and the formation of beliefs in turn.

**Retention Decision** Given the voter’s interim beliefs  $\tilde{\theta}_I^{e,y}(\cdot)$  about the incumbent ability, her net benefit of retaining the incumbent is:

$$\gamma_V \left( \tilde{\theta}_I^{e,y}(\cdot) - \theta_C \right) (q_H - q_L) + U(x_V; x_I, x_C)$$

She will thus choose to retain the incumbent i.f.f. :

$$\tilde{\theta}_I^{e,y}(\cdot) \geq \theta_C - \frac{U(x_V; x_I, x_C)}{\gamma_V (q_H - q_L)} = \bar{\theta}_C(x_V, x_I, x_C; \gamma_V, q_H, q_L),$$

where  $\bar{\theta}_C(\cdot)$  denotes the voter’s *belief threshold* for retention. To isolate attention to conflict between the saboteur and the voter, we henceforth restrict attention to the region of the parameter space within which the voter prefers to *retain* an incumbent known to be high ability, but *replace* an incumbent known to be low ability—that is, where  $\bar{\theta}_C(\cdot) \in (0, 1)$ .



**Belief Formation** We next calculate the voter's beliefs after each observable outcome. First consider when the saboteur does not engage in sabotage ( $e = 1$ ), so both success and failure are possible. After success, the voter's updated belief about the incumbent's ability is

$$\tilde{\theta}_I^{1,1}(\cdot) = \frac{\rho \cdot \theta_I q_H e_H + (1 - \rho) \cdot \theta_I q_H e_\emptyset}{\rho \cdot (\theta_I e_H q_H + (1 - \theta_I) e_L q_L) + (1 - \rho) \cdot q_\emptyset e_\emptyset}, \quad (2)$$

where  $q_\emptyset = \theta_I q_H + (1 - \theta_I) q_L$  is the expected probability of success when the saboteur is uninformed about the incumbent's ability ( $\eta = \emptyset$ ). After failure, her updated belief is

$$\tilde{\theta}_I^{1,0}(\cdot) = \frac{\rho \cdot \theta_I e_H (1 - q_H) + (1 - \rho) \cdot \theta_I (1 - q_H) e_\emptyset}{\rho \cdot (\theta_I e_H (1 - q_H) + (1 - \theta_I) e_L (1 - q_L)) + (1 - \rho) \cdot (1 - q_\emptyset) e_\emptyset} \quad (3)$$

Success causes the voter to revise her beliefs upward from what they would be after observing implementation alone, while failure causes her to update downward. Consequently,  $\tilde{\theta}_I^{1,1}(\cdot) > \tilde{\theta}_I^{1,0}(\cdot)$ , unless the saboteur's decision to implement the incumbent's policy has already perfectly signaled that the incumbent is high ability ( $e_H > 0 = e_L$  and  $\rho = 1$  so that  $\tilde{\theta}_I^{1,1}(\cdot) = \tilde{\theta}_I^{1,0}(\cdot) = 1$ ) or low ability ( $e_L > 0 = e_H$  and  $\rho = 1$  so that  $\tilde{\theta}_I^{1,1}(\cdot) = \tilde{\theta}_I^{1,0}(\cdot) = 0$ ).

Next consider if the saboteur engages in sabotage ( $e = 0$ ), after which failure is assured. After that failure, the voter's interim belief about the incumbent's ability is:

$$\tilde{\theta}_I^{0,0}(\cdot) = \frac{\rho \cdot \theta_I (1 - e_H) + (1 - \rho) \cdot \theta_I (1 - e_\emptyset)}{\rho \cdot (\theta_I (1 - e_H) + (1 - \theta_I) (1 - e_L)) + (1 - \rho) \cdot (1 - e_\emptyset)} \quad (4)$$

**Incumbent Popularity** Equilibrium turns out to depend crucially on what the voter's beliefs and retention decisions would be if sabotage and effort were themselves uninformative about the incumbent's ability. We therefore also specifically characterize these beliefs, and term the retention decisions that they lead to the incumbent's initial *popularity*.

After sabotage, the voter's beliefs would just remain at the prior  $\theta_I$  if sabotage were uninformative. The voter's beliefs after success and failure if effort were uninformative are  $\tilde{\theta}_I^{1,y}(1, 1, 1)$ ; that is, the beliefs characterized in equations 2-4 if the voter believed the saboteur to be *pooling on implementation* ( $e_L = e_H = e_\emptyset = 1$ ). Denoting these beliefs as  $\tilde{\theta}_I^y$ , we have:

$$\tilde{\theta}_I^1 = \frac{\theta_I q_H}{\theta_I q_H + (1 - \theta_I) q_L} \quad \text{and} \quad \tilde{\theta}_I^0 = \frac{\theta_I (1 - q_H)}{\theta_I (1 - q_H) + (1 - \theta_I) (1 - q_L)}$$

Clearly  $0 < \tilde{\theta}_I^0 < \theta_I < \tilde{\theta}_I^1 < 1$  (failure and success are imperfect “bad news” and “good news” about the incumbent’s ability, respectively). Using these beliefs, we now divide the incumbent’s initial popularity into four categories for the purposes of equilibrium analysis.

**Definition 1** *The incumbent is said to be*

(VU) **Very unpopular** *i.f.f.*  $\tilde{\theta}_I^0 < \theta_I < \tilde{\theta}_I^1 \leq \bar{\theta}_C(\cdot)$

(SU) **Somewhat unpopular** *i.f.f.*  $\tilde{\theta}_I^0 < \theta_I \leq \bar{\theta}_C(\cdot) < \tilde{\theta}_I^1$

(SP) **Somewhat popular** *i.f.f.*  $\tilde{\theta}_I^0 < \bar{\theta}_C(\cdot) \leq \theta_I < \tilde{\theta}_I^1$

(VP) **Very popular** *i.f.f.*  $\bar{\theta}_C(\cdot) \leq \tilde{\theta}_I^0 < \theta_I < \tilde{\theta}_I^1$

A popular incumbent is one who would be retained in the absence of new information (either from the saboteur’s observed effort decisions, the policy outcome, or both), while an unpopular incumbent is one who would be replaced. The distinction between a “very” and “somewhat” popular or unpopular incumbent is based on what the voter would do after observing success or failure (but inferring nothing from the absence of sabotage alone); she would follow her prior for a “very” popular or “very” unpopular incumbent regardless of the outcome, but base her retention decisions on observed success or failure for a “somewhat” popular or “somewhat” unpopular incumbent. Figure 1 provides a visual summary of the definition of popularity.

## 4 An uninformed saboteur ( $\rho = 0$ )

We first consider the variant of the model in which it is common knowledge that the saboteur is no better informed than the voter ( $\rho = 0$ ), so that the voter infers nothing *directly* from the saboteur’s decision. She thus follows her prior  $\theta_I$  if she observes sabotage since it suppresses the revelation of additional information about ability via outcomes; otherwise she updates her beliefs to  $\tilde{\theta}_I^y$  based on success or failure.

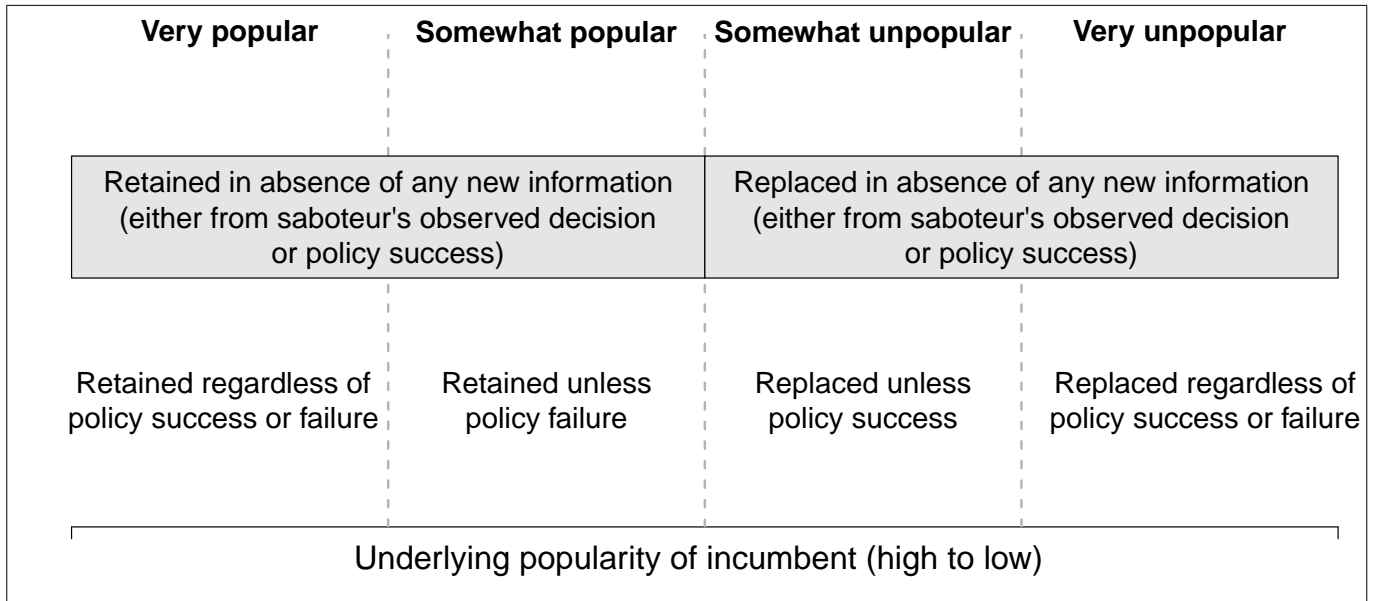


Figure 1: The popularity of the incumbent and its relationship to retention.

**Proposition 1** *Suppose that the saboteur is uninformed about the incumbent's ability. Then there is a unique sequential equilibrium in which the saboteur sabotages if and only if the incumbent is somewhat unpopular.*<sup>8</sup>

Figure 2 summarizes the equilibrium; the top panel depicts the probability of sabotage, while the bottom panel depicts the probability the incumbent is retained as a function of both the sabotage decision and policy success or failure. In both panels, the horizontal axis captures the probability  $\theta_C$  that the challenger is high ability—varying this parameter over  $[0, 1]$  generates the four popularity regions. In the bottom panel, the dashed (red) lines depict the probability that a low-ability incumbent is sabotaged, while the solid (green) lines depict the probability that a high-ability incumbent is sabotaged.

The calculus of an uninformed saboteur is straightforward. He can allow the incumbent's policy to proceed naturally, which will result in success with a probability associated with the incumbent's ability. Alternatively, he can sabotage, which suppresses information revelation about the incumbent's ability, but also destroys the possibility of a successful outcome. If the incumbent is very popular or very unpopular there is no benefit to sabotage; the voter's

<sup>8</sup>Sequential equilibrium (Kreps and Wilson 1982) is necessary to ensure that the voter does not “infer” something from sabotage off the equilibrium path that an uninformed saboteur cannot know.

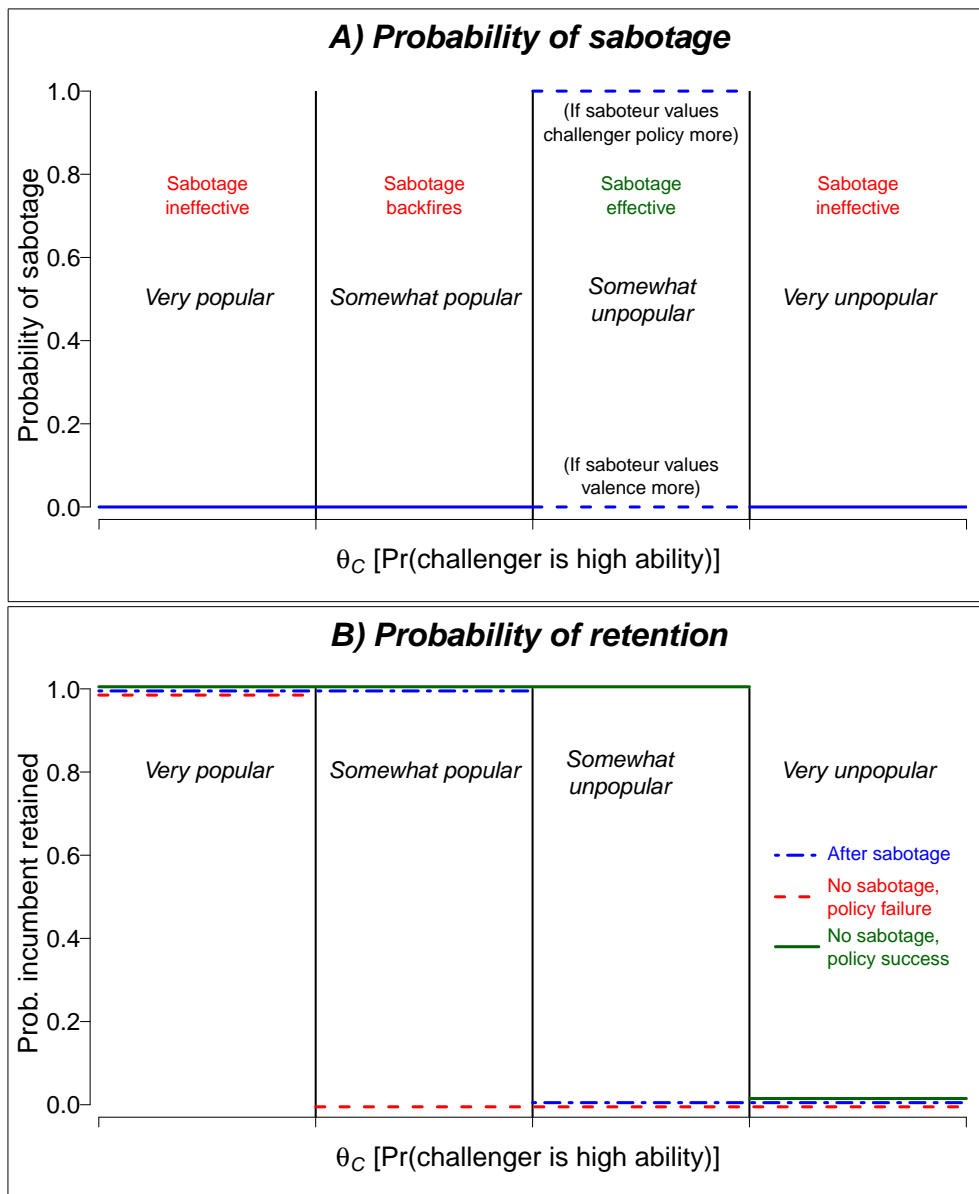


Figure 2: Summary of equilibrium results when the saboteur is **uninformed** about the ability of the incumbent. The top panel depicts the probability of sabotage across the four popularity regions, while the bottom panels depict the probability the incumbent is retained.

decision will be unaffected, so it simply results in foregone success. When the incumbent is somewhat popular, sabotage actually *backfires*; not only would the voter retain the incumbent and the saboteur would eliminate the chance of success, but sabotage eliminates the possibility that voter would learn through failure that she wishes to *replace* the incumbent. Only when the incumbent is *somewhat unpopular* does sabotage make sense. In this case, if the saboteur implements the policy, a success would lead the voter to retain. Sabotage thus deprives the voter of the opportunity to learn through success that she would actually prefer

to retain an incumbent who she would have otherwise replaced, and is therefore profitable if the saboteur's net ideological benefit for the challenger is sufficiently high.

## 5 An informed saboteur ( $\rho = 1$ )

When the saboteur is known to be privately informed about the incumbent's ability, the effect of sabotage on the voter's beliefs is more complicated. Sabotage suppresses the revelation of information about the incumbent's ability via *policy outcomes*. However, since the voter understands that the decision to sabotage is strategic, sabotage may *itself* signal that the incumbent is high ability if the act represents the saboteur's attempt to conceal this fact. Similarly, the calculus of a privately informed saboteur is more complicated because he does not have an unambiguously greater incentive to sabotage one type of incumbent or the other. Instead, there are two competing forces, and which one dominates depends on exactly how the voter uses policy outcomes in her retention decisions.

The first force is the saboteur's intrinsic value for success. Since the policy of a high-ability incumbent is more likely to generate a success than a low-ability one, more utility is lost (in expectation) when the policy of a high-ability incumbent is sabotaged. This force creates a greater willingness to sabotage a low-ability incumbent versus a high-ability one. In the preceding analysis this property is manifested in  $\bar{\Delta}_H(\cdot) > \bar{\Delta}_L(\cdot)$ ; that is, a higher electoral impact of sabotage is necessary to induce sabotage of a high-ability incumbent versus a low-ability one.

The second force is the potentially greater electoral competitiveness of high-ability incumbents. If their policies are implemented, high ability incumbents are more likely to succeed by virtue of their greater ability. If the voter is strongly basing her retention decision on success and failure ( $\pi_1^1 - \pi_1^0$  is large), then high-ability incumbents are therefore also more likely to be retained if not sabotaged. This force pushes in the direction of a greater willingness to sabotage a high ability incumbent versus a low-ability one, as sabotage is more likely to block a successful outcome that would improve the incumbent's electoral fortunes.

In equilibrium, what a rational voter infers from the decision to sabotage or not depends

on which of these two forces dominates. As we will show, the presence of these competing forces can both *increase* or *decrease* the amount of sabotage that occurs in equilibrium relative to when the saboteur is uninformed. In addition, it is possible for a rational saboteur to engage in sabotage even though it *improves* the incumbent’s reputation.

Determining what a rational voter should infer from sabotage or effort also requires a way of determining the voter’s beliefs when she expects one action from the saboteur (e.g. always sabotage regardless of the incumbent’s ability), but instead sees the other (e.g. effort). For this we apply an equilibrium refinement in the spirit of D1 (Cho and Kreps 1987)—henceforth called simply D1. This effectively states that “off the equilibrium path,” the voter should believe that the incumbent is of an ability that would have induced the saboteur to take the unexpected action for the largest set of “reasonable” responses by the voter.<sup>9</sup>

## 5.1 A Somewhat Unpopular Incumbent

We first discuss what we see as the most interesting case, that of a somewhat unpopular incumbent. Such an incumbent will be replaced unless a policy success occurs that improves her reputation; recall that when the saboteur is uninformed, this induces him to sabotage. When the saboteur is informed, however, sabotage may backfire and improve the incumbent’s reputation if it signals that the incumbent is high ability. Can this effect induce a privately informed saboteur to refrain from sabotage in equilibrium that he would otherwise undertake if uninformed? The answer is yes:

**Proposition 2** *If the saboteur is informed and the incumbent is somewhat unpopular, then the following equilibrium satisfies D1: the saboteur never sabotages, and the voter only replaces after seeing both implementation and failure.*

When the voter *expects* the saboteur to refrain from sabotage and the incumbent is somewhat unpopular (or somewhat popular), policy outcomes maximally influence her retention

---

<sup>9</sup>Determining the set of “reasonable” off-path responses for the voter requires a modification to D1 since nature has an intervening move after the saboteur’s decision. See the Appendix for details.

decision—she will retain the incumbent if and only if the policy succeeds ( $\pi_1^1 - \pi_1^0 = 1$ ). This electoral behavior maximizes the saboteur’s incentive to sabotage a high-ability incumbent, which in turn leads the voter to infer from *unexpected* sabotage that the incumbent is high ability and retain him, which then induces the saboteur to refrain from it.

We next ask whether it is also possible for a privately informed saboteur to always sabotage in equilibrium, *even though* the voter knows that she is privately informed about the incumbent’s ability. Somewhat surprisingly, the answer is also yes.

**Proposition 3** *If the saboteur is informed and the incumbent is somewhat unpopular, then the following equilibrium satisfies D1: the saboteur always sabotages, the voter always replaces the incumbent, and should the voter unexpectedly see effort she retains.*

When the voter *expects* sabotage and the incumbent is somewhat unpopular, it will lead her to replace—not because she infers anything from sabotage itself, but precisely because she does not. However, the effectiveness of sabotage on both low- and high-ability incumbents, combined with the greater intrinsic cost of sabotaging high-ability incumbents, actually makes it costlier to sabotage a high-ability incumbent. Consequently, should the saboteur unexpectedly decline to sabotage, the voter will infer that the incumbent is high ability and retain regardless of the policy, which in turn induces the saboteur to sabotage.

The preceding analysis illustrates the complexity of predicting what a rational voter should infer from sabotage or its absence, and thus what the saboteur will do—what the voter will think depends strongly on what she expects. However, both of the preceding equilibria have the undesirable property that one action is “off the equilibrium path,” which requires a criteria (D1) for determining what the voter should believe if she sees an unexpected action. As it turns out, there exists an equilibrium in which both sabotage and its absence occur:

**Proposition 4** *If the saboteur is informed and the incumbent is somewhat unpopular, then the following is an equilibrium:*

- The saboteur’s probability of implementation for each type of incumbent is

$$0 < e_H = \left( \frac{q_L}{q_H - q_L} \right) \left( \frac{\bar{\theta}_C(\cdot) - \theta_I}{\theta_I(1 - \bar{\theta}_C(\cdot))} \right) < e_L = \left( \frac{q_H}{q_H - q_L} \right) \left( \frac{\bar{\theta}_C(\cdot) - \theta_I}{\bar{\theta}_C(\cdot)(1 - \theta_I)} \right) < 1$$

- The voter’s probabilities of retaining after failure, sabotage, and success are

$$\pi_1^0 = 0 < \pi_0^0 = \frac{q_L \bar{\Delta}_H(\cdot) - q_H \bar{\Delta}_L(\cdot)}{q_H - q_L} < \pi_1^1 = \frac{\bar{\Delta}_H(\cdot) - \bar{\Delta}_L(\cdot)}{q_H - q_L} < 1$$

We label this equilibrium the “sometimes sabotage” equilibrium; its structure is as follows. First, the saboteur sometimes sabotages *both* high- and low-ability incumbents; somewhat surprisingly, however, he is actually *more likely* to sabotage a high-ability one! Sabotage thus perversely *improves* the incumbent’s reputation, leading the voter to sometimes retain them. Conversely, the absence of sabotage harms the incumbent’s reputation, but not so much that a policy success cannot overcome it. An incumbent who succeeds is retained with a higher probability than an incumbent who is sabotaged, but an incumbent who fails is always replaced.

In the “sometimes sabotage” equilibrium, both forces that potentially influence the saboteur’s incentive to sabotage operate. The saboteur’s intrinsic preferences for success makes sabotaging a high-ability incumbent intrinsically more costly. Simultaneously, the voter’s use of outcomes in her retention decisions makes sabotaging a high-ability incumbent more electorally damaging. In equilibrium, these forces exactly balance each other out, leading the saboteur to sometimes sabotage both types of incumbents.

We summarize this equilibrium in Figure 3, which parallels Figure 2 except that it summarizes the results when the saboteur is informed of the incumbent’s ability. For purposes of comparison, the figure also present the results with an informed saboteur from the other three popularity regions, which are discussed next. Notably, as compared to when the saboteur is uninformed, the incumbent’s electoral prospects are not as bleak after sabotage; sabotage by an uninformed saboteur always leads the voter to replace, but sabotage by an informed saboteur only sometimes does ( $\pi_0^0 > 0$ ) due to the reputational benefit that sabotage brings. Conversely, as compared to when the saboteur is uninformed the incumbent’s



electoral prospects are not as secure after success; success when the saboteur is uninformed always leads to reelection, but when he is informed it only sometimes does ( $\pi_1^1 < 0$ ) due to the reputational harm that sabotage's *absence* inflicts. The preceding two observations yield the following corollary to Proposition 4.

**Corollary 1** *For either incumbent type  $\lambda_I \in \{L, H\}$ , the equilibrium electoral impact of sabotage  $\Delta_{\lambda_I}(\cdot) = q_{\lambda_I}\pi_1^1 - \pi_0^0$  is larger in the model with an uninformed saboteur ( $\rho = 0$ ) than with an informed saboteur ( $\rho = 1$ ) in the “sometimes sabotage” equilibrium.*

### Welfare and Comparative Statics

Having characterized three potential equilibria, we now ask whether it is possible to select one using a welfare criteria—is one equilibrium superior for both players? Unfortunately, the answer is no; the voter is better off in equilibria with less sabotage, while the saboteur is better off in equilibria with more.

**Proposition 5** *Pooling on implementation is best for the voter and worst for the saboteur. Pooling on sabotage is best for the saboteur and worst for the voter. Sometimes sabotaging is intermediate for both players.*

An interesting implication of Proposition 5 is that the saboteur does not benefit from (and can even be harmed by) having superior information, as it can cause sabotage to backfire.

Despite the absence of a rigorous criterion for selection, the equilibrium in which sabotage sometimes occurs has a certain empirical plausibility, as it is the only one in which both sabotage and its absence occur on the equilibrium path. It is also the most interesting, as it clearly illustrates how all of the following are possible: a rational voter can understand the saboteur's greater incentive to sabotage a high-ability incumbent; the voter can respond to it by sometimes reelecting such an incumbent; yet the saboteur may nevertheless sometimes pursue sabotage.

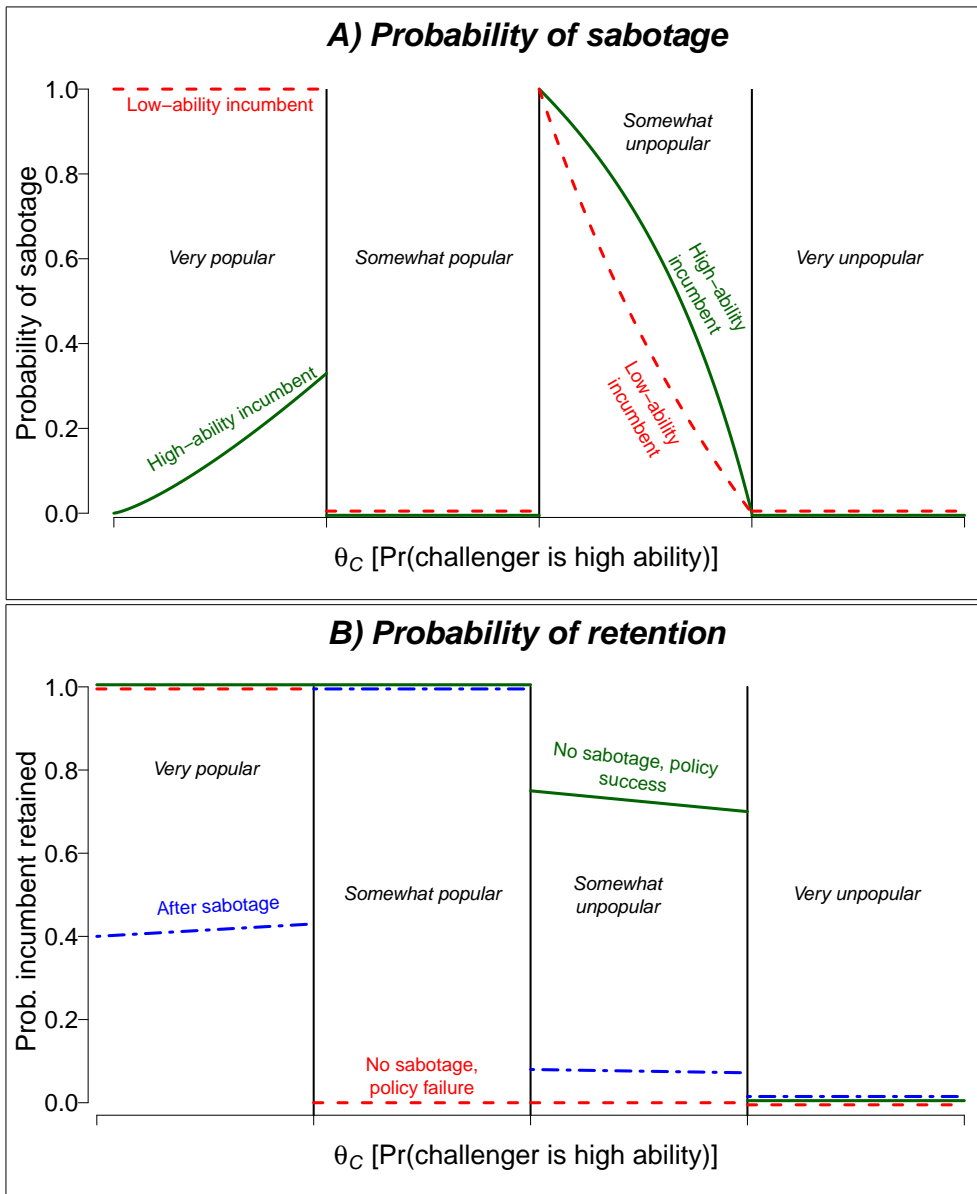


Figure 3: Summary of equilibrium results when the saboteur is *informed* about the ability of the incumbent—specifically, in the “sometimes sabotage” equilibrium (see text for details). The top panel depicts the probability of sabotage across the four popularity regions, while the bottom panels depict the probability the incumbent is retained.

We conclude this section by examining comparative statics in this equilibrium (behavior is invariant to the underlying model parameters in the two pooling equilibria). We first consider the saboteur’s probability of sabotaging.

**Proposition 6** *When the incumbent is somewhat unpopular, the probability  $1 - e_{\lambda_I}$  that the saboteur sabotages each type of incumbent is:*

- decreasing in the challenger’s reputation,  $\theta_C$ , and in the voter’s value for success  $\gamma_V$

- increasing in the voter’s net ideological benefit for the incumbent  $U(x_V; x_I, x_C)$  and the incumbent’s reputation  $\theta_I$

When the incumbent is somewhat unpopular, the key determinant of the likelihood that each type of incumbent is sabotaged is the need to keep the voter indifferent over retaining the incumbent in the face of both sabotage *and* policy success. If the voter becomes more inclined *ex ante* to replace the incumbent *ceteris paribus* (higher  $\theta_C$  or  $\gamma_V$ , lower  $U(x_V; x_I, x_C)$  or  $\theta_I$ ), then sabotage must become stronger “good news” about the incumbent to lead the voter to sometimes retain after sabotage. Simultaneously, implementation must become weaker “bad news” about the incumbent to make the voter willing to sometimes retain after success. Thus, the presence of sabotage must become more informative about the incumbent’s ability, but its absence less informative, so that the total amount of information communicated by the saboteur’s behavior is constant. This can only be accomplished by having the probability of sabotaging both types of incumbent decrease.

We next examine the voter’s retention probabilities.

**Proposition 7** *When the incumbent is somewhat unpopular, the voter’s probability of retaining the incumbent after both success  $\pi_1^1$  and after sabotage  $\pi_0$ , as well as the difference between them  $\pi_1^1 - \pi_0$ , is decreasing in the saboteur’s relative benefit  $-\frac{U(x_S; x_I, x_C)}{\gamma_S}$  for the challenger policy, his weight on the future  $\delta$ , and the challenger’s reputation  $\theta_C$ .*

The key determinant of the probabilities that the incumbent is retained following both success and sabotage is the need to keep the saboteur indifferent over sabotaging both types of incumbents. In the “sometimes sabotage” equilibrium, the impact of effort on the probability an incumbent of each type is retained is  $\Delta_{\lambda_i} = q_{\lambda_i} \pi_1^1 - \pi_0$ . Recalling that  $\bar{\Delta}_{\lambda_I}$  is the impact probability that makes the saboteur indifferent to sabotaging an incumbent of ability  $\lambda_I$ , equilibrium requires that  $\bar{\Delta}_{\lambda_I} = q_{\lambda_I} \pi_1^1 - \pi_0 \forall \lambda_i$ .

To understand the comparative statics, we first note (and prove in the Appendix) that  $\bar{\Delta}_{\lambda_I}$ ,  $\bar{\Delta}_H - \bar{\Delta}_L$ , and  $\bar{\Delta}_H - \frac{q_H}{q_L} \bar{\Delta}_L$  are all decreasing in  $-\frac{U(x_S; x_I, x_C)}{\gamma_S}$ ,  $\delta$ , and  $\theta_C$ . In words, as

the saboteur’s weight on ideology, the saboteur’s weight on the future, or the challenger’s reputation increase, the saboteur’s willingness to sabotage both types of incumbents both increase and become more similar. These observations imply the desired comparative statics. First, the voter’s probability of retaining after success  $\pi_1^1$  determines how *different* is the electoral impact of effort for a high- versus low-ability incumbent, since they have different likelihoods of succeeding. When the difference  $\bar{\Delta}_H - \bar{\Delta}_L$  in the thresholds that trigger sabotage for each type decrease, so too must  $\pi_1^1$ . For a similar reason, the probability of retaining after sabotage  $\pi_0$  must decrease in  $\bar{\Delta}_H - \frac{q_H}{q_L} \bar{\Delta}_L$ . Finally, the difference  $\pi_1^1 - \pi_0$  in the probability of retaining after success and sabotage effectively captures the electoral impact of effort for both incumbent types; so as both thresholds  $(\bar{\Delta}_H, \bar{\Delta}_L)$  for sabotage decrease, so too must  $\pi_1^1 - \pi_0$ .

## 5.2 A Very Popular Incumbent

We next transition to the case of a very popular incumbent. A key finding from the case of a somewhat unpopular incumbent is that being more informed about the incumbent’s ability may cause the saboteur to engage in sabotage less often. We now show that when the incumbent is very popular, the opposite is true: an informed saboteur will sabotage *more* than an uninformed one.

Recall from Section 4 that an *uninformed* saboteur never sabotages a very popular incumbent because nothing can be accomplished from doing so—the voter will simply retain them absent new information. For the same reason, it cannot be an equilibrium for an informed saboteur to *always* (that is, regardless of the incumbent’s ability) sabotage a very popular incumbent—the voter will neither observe outcomes nor infer anything from sabotage, and will thus retain the incumbent for sure.

The logic breaks down, however, when considering whether it is an equilibrium for an informed saboteur to *never* sabotage a very popular incumbent. When the voter knows that the saboteur is informed about the incumbent’s ability, the unexpected presence of sabotage itself contains information about that ability. The effectiveness and incidence of sabotage

thus hinge on a simple question—what will the voter infer about the incumbent’s ability in the face of *unexpected* sabotage? When the incumbent is very popular, the answer is simple: the voter will infer that the incumbent is low ability and replace him. Somewhat counterintuitively, the reason is that the saboteur also intrinsically values policy success. If the incumbent is so popular *ex ante* that he will be retained even after policy failure, then there is *no greater electoral benefit* to sabotaging a high-ability incumbent than a low-ability one—absent sabotage, the former will succeed with a higher probability than the latter, but both will be retained regardless. However, it remains intrinsically costlier to sabotage a high-ability incumbent. The voter will therefore infer that a very popular incumbent who is unexpectedly sabotaged is definitely low ability and replace her, incentivizing the saboteur to indeed sabotage, and causing such an equilibrium to unravel.

It turns out that when the incumbent is very popular, there is a unique equilibrium that satisfies D1; it is partially separating, and takes the following form. First, sabotage must sometimes occur and harm the incumbent’s electoral prospects, and so must credibly communicate *some* negative information about the incumbent’s ability. However, it cannot perfectly communicate that the incumbent is low ability; if it did, then sabotage would cause the incumbent to be replaced for sure, and the saboteur would always want to sabotage regardless of the incumbent’s ability. Thus, in equilibrium the saboteur must *always* sabotage a low-ability incumbent ( $e_L = 0$ ), and *sometimes* sabotage a high-ability one ( $e_H > 0$ ). With this strategy, the absence of sabotage perfectly reveals that the incumbent is high ability and ensures reelection, regardless of whether the incumbent’s policy succeeds or fails. The presence of sabotage, in contrast, credibly but *imperfectly* reveals that the incumbent is low ability, triggering replacement with a strictly positive probability. Formally, the equilibrium is as follows.

**Proposition 8** *Suppose that the saboteur is informed and the incumbent is very popular. Then there is a unique equilibrium satisfying refinement (D1) that takes the following form.*

- *The saboteur always sabotages a low-ability incumbent’s policy ( $e_L = 0$ ) and implements*

a high-ability incumbent's policy with probability  $e_H = \frac{\theta_I - \bar{\theta}_C(\cdot)}{\theta_I(1 - \bar{\theta}_C(\cdot))}$

- The voter always retains the incumbent absent sabotage regardless of the outcome ( $\pi_1^0 = \pi_1^1 = 1$ ), and retains after sabotage with an interior probability equal to  $\pi_0^0 = 1 - \bar{\Delta}_H(\cdot)$

Equilibrium thus exhibits a great deal of sabotage that would not occur if the saboteur were uninformed, and sabotage definitively harms the incumbent's electoral prospects.

**Comparative Statics** We first consider the saboteur's probability of sabotaging.

**Corollary 2** *When the incumbent is very popular, the saboteur always sabotages a low-ability incumbent. The probability he sabotages a high-ability incumbent is:*

- increasing in the challenger's reputation,  $\theta_C$ , and in the importance to the voter  $\gamma_V (q_H - q_L)$  of having a high-ability incumbent.
- decreasing in the voter's net ideological preference for the incumbent  $U(x_V; x_I, x_C)$  and the incumbent's reputation  $\theta_I$

When the incumbent is very popular, the key determinant of the likelihood of sabotage is the need to keep the voter indifferent over retaining the incumbent in the face of sabotage. If the voter's desire to retain the incumbent after sabotage increases *ceteris paribus* (higher  $U(x_V; x_I, x_C)$  or  $\theta_I$ ), then sabotage must become a more credible signal that the incumbent is low ability to maintain indifference, and thus the saboteur must sabotage a high-ability incumbent less often. Conversely, if the voter's desire to reelect the incumbent after sabotage decreases *ceteris paribus* (higher  $\theta_C$  or  $\gamma_V (q_H - q_L)$ ), then sabotage must become a less credible signal that the incumbent is low ability to maintain indifference, and thus the saboteur must sabotage a high-ability incumbent more often.

We last examine the voter's likelihood of retaining a sabotaged incumbent.

**Corollary 3** *When the incumbent is very popular, the voter's probability of retaining the incumbent after sabotage is:*

- increasing in the saboteur’s relative value  $-\frac{U(x_S;x_I,x_C)}{\gamma_S}$  for the challenger, his weight on the future  $\delta$ , the challenger’s reputation  $\theta_C$ , and the probability  $q_L$  that a low-ability politician succeeds
- decreasing in the probability  $q_H$  that a high-ability politician succeeds.

When the incumbent is very popular, what determines the likelihood that the incumbent is retained after sabotage is the need to keep the saboteur indifferent over sabotaging a high-ability incumbent. The higher is the likelihood that the incumbent is *still* retained despite sabotage, the lower is the saboteur’s incentive to engage in it. Thus, if the saboteur’s electoral incentive to sabotage a high-ability incumbent goes up (due to a greater net ideological benefit for the challenger  $-U_S(\cdot)$ , a greater weight on the future  $\delta$ , a challenger likelier to be high ability  $\theta_C$ , or a decreased importance of selecting high-ability politicians  $q_L$ ), then the voter’s likelihood of retaining the incumbent post-sabotage must increase to maintain indifference. Conversely, if the saboteur becomes less willing to sabotage because the importance of selecting high-ability incumbents goes up (higher  $q_H$  or  $\gamma_S$ ), the likelihood of retaining the incumbent post-sabotage must decrease to maintain indifference.

### 5.3 Very Unpopular and Somewhat Popular Incumbents

We last consider the cases of very unpopular and somewhat popular incumbents. As it turns out, in these cases the behavior of an informed saboteur is exactly the same as that of an uninformed one: he never sabotages.

First consider a very unpopular incumbent, and recall the reason that an uninformed saboteur never sabotages— he will get his desired electoral outcome either way. By the same logic, when the saboteur is informed it remains an equilibrium to never sabotage regardless of the incumbent’s ability; since the incumbent already has no electoral prospects, sabotage cannot make them any worse.<sup>10</sup>

---

<sup>10</sup>There are also two additional equilibria that satisfy D1—one in which the saboteur sometimes sabotages a low-ability incumbent, and one in which he always sabotages both types of incumbents. However, because both of these equilibria are Pareto-dominated by

**Proposition 9** *If the saboteur is informed and the incumbent is very unpopular, then there is a unique Pareto-dominant equilibrium among those satisfying refinement (D1) in which the saboteur never sabotages and the incumbent is always replaced.*

We last consider a somewhat popular incumbent. Recall that an uninformed saboteur also never sabotages a somewhat popular incumbent, because it will simply prevent the voter from learning via failure that she wishes to replace them. It turns out that never sabotaging remains an equilibrium when the saboteur is informed, but for somewhat more subtle reasons. Similar to the somewhat unpopular case, when the incumbent is somewhat popular the voter will infer from *unexpected* sabotage that the incumbent is *definitely* high ability and should be retained, due to the saboteur’s greater electoral incentive to sabotage a high-ability incumbent. As a result, the saboteur knows that sabotage would backfire and avoids it.<sup>11</sup>

**Proposition 10** *Suppose that the saboteur is informed and the incumbent is somewhat popular. Then there is a unique Pareto-dominant equilibrium among those satisfying refinement (D1) in which the saboteur never sabotages.*

## 6 Discussion and Conclusion

We have presented a model of policy sabotage, in which the potential for a saboteur to intervene in policy implementation complicates a voter’s ability to select a politician who will perform best in office. While many of our results are intuitive, we showed that the interaction of the possibility of sabotage and concerns over policy combine to create complicated incentives for a would-be saboteur, and a challenging informational environment for the equilibrium in which the saboteur never sabotages, we omit their consideration from the main text.

---

<sup>11</sup>There are again two additional equilibria satisfying D1—one in which the saboteur always sabotages a low-ability incumbent and sometimes sabotages a high-ability one, and another in which he sometimes sabotages both types. However, because both of these equilibria are Pareto-dominated by the equilibrium in which the saboteur never sabotages, we again omit consideration from the main text.



voters. We motivated our inquiry with the following question: how can *observable* sabotage be rational for an out-party if the voter understands why such sabotage is occurring? Our model provides one answer. An opposition party does not sabotage because it thinks it will harm an incumbent's reputation with rational voters. Rather, it sabotages *despite* the fact that sabotage will improve the incumbent's reputation with rational voters, fearing that the absence of sabotage and a policy success will improve that reputation even more.

More generally, while we have framed our model somewhat narrowly around the notion of policy sabotage, the model also has larger implications about the tradeoffs between policy making and electoral considerations, particularly for the minority- or out-party in a separated powers system. On this point, the results in our model line up nicely with the arguments in Frances Lee's (2016) book, *Insecure Majorities*. Lee's central argument is that the relatively recent trend in American politics towards highly competitive national elections, in which the Democratic and Republican parties always have a decent chance of moving from minority status in Congress to majority status in any given election (and vice versa), has fundamentally changed the calculus of the minority party. When, for example, Democrats dominated congressional elections between the 1950s and the 1990s, the possibility of being in the majority was not something Republicans had to take seriously, and thus they could focus more on straightforward policy victories, whenever available. Under the current reality of high-stakes partisan competition over Congress, however, the minority party is often incentivized to forsake immediate policy gains. This is because it hopes that blocking legislation favored by the majority party will make it more likely that the minority party will win (or do better than expected) in the next election. Lee argues that such incentives lead to the proliferation of "messaging" legislation, in which the goal of a party is not to put forward proposals that have a serious chance of enactment, but rather to curry favor with voters.

It's easy to see how policy sabotage fits as a tool for a minority- or out-party to use in the current political environment. But our model provides a micro-founded rationalization for why such tactics can produce electoral benefits, even when they are observable by voters.

One way to understand the rise of messaging politics is that voters are not very well informed about politics or policy, and that messaging legislation is simply a means to get voters to irrationally focus on optics over substance. What our model shows that even if voters are fully rational, it would still be rational for parties to engage in tactics like sabotage or messaging legislation, because failure to do so may hurt their chances of winning elections even more. Furthermore, our model predicts that under certain conditions, it would also be rational to engage in sabotage or messaging politics even when the incumbent party or president is very popular.<sup>12</sup>

Moving beyond the politics of sabotage by an opposing minority, an interesting possibility we leave for future analysis is that a saboteur who is ideologically *aligned* with the incumbent may also wish to engage in sabotage. Specifically, an ideologically-aligned saboteur may sabotage to prevent a moderately-strong incumbent from undertaking a “policy gamble” that, if unsuccessful, would damage her reputation sufficiently to result in electoral turnover (Dewan and Hortala-Vallve 2017). Moreover, a rational voter might expect an aligned saboteur to be most willing to shield the incumbent from accountability via sabotage precisely when he lacks confidence in the incumbent’s policy. Once a broader set of ideological preferences for the saboteur are considered, or even the possibility of competing saboteurs with differing ideological allegiances, then it becomes clear that politics of sabotage, and the corresponding informational problem of voters, may be complex indeed. Such questions are examined in contemporaneous work by Kang and Park (2021), which analyzes the politics of information production in a related but distinct model that lacks a signaling component.

Finally, while we chose to set our model within the broader literature on democratic accountability, other paths are available. For instance, certain actors—such as bureaucrats deep within the bowels of the federal bureaucracy—may be able to engage in sabotage without it being immediately observable to voters. This lack of detectability could both

---

<sup>12</sup>We thank an anonymous reviewer for highlighting this connection between Lee’s arguments and our model.

increase or decrease the amount of sabotage that occurs, as compared to the model we have analyzed.<sup>13</sup> Sabotage may also have different effects from the one we have studied—for instance, it may change the status quo of a policy and/or the reversion point, thereby opening up opportunities for future bargaining. A “pivotal politics”-style model could pursue this path. Alternatively, where we modeled a single voter, sabotage may please some voters at the expense of others. Thus, a model with heterogeneous voters could produce additional insights. Finally, while we have focused on sabotage within the context of horizontally shared powers, the logic of our model could easily be extended to examine the incentives for sabotage in a system of federalism where local actors oppose national policies (Bulman-Pozen and Gerken 2008).

## References

- Arceneaux, Kevin. 2005. “Does Federalism Weaken Democratic Representation in the United States?” *Publius: The Journal of Federalism* 35(2):297–311.
- Arceneaux, Kevin. 2006. “The Federal Face of Voting: Are Elected Officials Held Accountable for the Functions Relevant to their Office?” *Political Psychology* 27(5):731–754.
- Ashworth, Scott. 2012. “Electoral Accountability: Recent Theoretical and Empirical Work.” *Annual Review of Political Science* 15:183–201.
- Ashworth, Scott and Ethan Bueno De Mesquita. 2008. “Electoral Selection, Strategic Challenger Entry, and the Incumbency Advantage.” *Journal of Politics* 70(4):1006–1025.
- Ashworth, Scott and Ethan Bueno De Mesquita. 2014. “Is Voter Competence Good for Voters?: Information, Rationality, and Democratic Performance.” *American Political Science Review* 108(3):565–587.

---

<sup>13</sup> Under this scenario, the voter would only observe policy success or failure. Success would imply that the saboteur implemented the policy, but failure could occur either via sabotage or a “true” failure. As it turns out, the results when sabotage is unobservable are either more obvious or less interesting than the results from the variants we have presented. Unobservability has the obvious effect of increasing the saboteur’s incentive to engage in sabotage. However, it also makes it more difficult to credibly “signal” that the incumbent is low ability via sabotage because the signal is mixed up with signals of failure due to the incumbent’s ability. These two effects mean that making sabotage unobservable can both increase and decrease the equilibrium amount of sabotage, depending on the region. Perhaps most importantly, with unobservable sabotage, the saboteur will always be (weakly) more likely to sabotage a low-ability incumbent than a high-ability one, and thus in equilibrium policy success (failure) will always be a signal that the incumbent is high (low) ability.

- Ashworth, Scott and Ethan Bueno de Mesquita. 2017. “Unified versus Divided Political Authority.” *The Journal of Politics* 79(4):1372–1385.
- Buisseret, Peter. 2016. ““Together or Apart”? On Joint versus Separate Electoral Accountability.” *The Journal of Politics* 78(2):542–556.
- Bulman-Pozen, Jessica and Heather K Gerken. 2008. “Uncooperative Federalism.” *Yale Law Journal* 118(7):1256–1310.
- Cameron, Charles M. 2000. *Veto Bargaining: Presidents and the Politics of Negative Power*. Cambridge: Cambridge University Press.
- Canes-Wrone, Brandice, Michael C Herron and Kenneth W Shotts. 2001. “Leadership and Pandering: A Theory of Executive Policymaking.” *American Journal of Political Science* 45(3):532–550.
- Cho, In-Koo and David M Kreps. 1987. “Signaling Games and Stable Equilibria.” *The Quarterly Journal of Economics* 102(2):179–221.
- Dewan, Torun and Rafael Hortala-Vallve. 2017. “Electoral Competition, Control and Learning.” *British Journal of Political Science* 49(3):1–17.
- Downs, George W and David M Roche. 1994. “Conflict, Agency, and Gambling for resurrection: The Principal-Agent Problem Goes to War.” *American Journal of Political Science* pp. 362–380.
- Epstein, David and Peter Zemsky. 1995. “Money Talks: Deterring Quality Challengers in Congressional Elections.” *American Political Science Review* 89(2):295–308.
- Fearon, James D. 1999. Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance. In *Democracy, Accountability, and Representation*, ed. Susan Stokes Adam Przeworski and Bernard Manin. Cambridge: Cambridge University Press.
- Ferejohn, John. 1986. “Incumbent Performance and Electoral Control.” *Public Choice* 50(1/3):5–25.
- Fong, Christian and Keith Krehbiel. 2018. “Limited Obstruction.” *American Political Science Review* 112(1):1–14.
- Gieczewski, Germán and Christopher Li. 2021. “Dynamic Policy Sabotage.” *American Journal of Political Science* .
- Gordon, Sanford C and Gregory A Huber. 2002. “Citizen Oversight and the Electoral Incentives of Criminal Prosecutors.” *American Journal of Political Science* 46(2):334–351.
- Gordon, Sanford C, Gregory A Huber and Dimitri Landa. 2007. “Challenger Entry and Voter Learning.” *American Political Science Review* 101(2):303–320.

- Groseclose, Tim and Nolan McCarty. 2001. "The Politics of Blame: Bargaining before an Audience." *American Journal of Political Science* pp. 100–119.
- Healy, Andrew, Alexander G Kuo and Neil Malhotra. 2014. "Partisan Bias in Blame Attribution: When Does it Occur?" *Journal of Experimental Political Science* 1(2):144–158.
- Healy, Andrew and Neil Malhotra. 2013. "Retrospective Voting Reconsidered." *Annual Review of Political Science* 16:285–306.
- Hirsch, Alexander V and Kenneth W Shotts. 2012. "Policy-Specific Information and Informal Agenda Power." *American Journal of Political Science* 56(1):67–83.
- Hirsch, Alexander V. and Kenneth W. Shotts. 2015. "Competitive Policy Development." *American Economic Review* 105(4):1646–64.
- Hirsch, Alexander V. and Kenneth W. Shotts. 2018. "Policy-Development Monopolies: Adverse Consequences and Institutional Responses." *Journal of Politics* 80(4):1339–54.
- Hitt, Matthew P., Craig Volden and Alan E. Wiseman. 2017. "Spatial Models of Legislative Effectiveness." *American Journal of Political Science* 61(3):575–590.
- Izzo, Federica. 2020. "With Friends Like These, Who Needs Enemies?" London School of Economics working paper. Available at <https://www.federicaizzo.com/pdf/WFLT10K.pdf>.
- Kang, Myunghoon and Ju Yeon Park. 2021. "Who Wants to Rock the Boat? Partisan Competition over the Clarity of Responsibility." Bilkent University working paper.
- Kreps, David M and Robert Wilson. 1982. "Sequential Equilibria." *Econometrica* 50:863–94.
- Lax, Jeffrey R. and Charles M. Cameron. 2007. "Bargaining and Opinion Assignment on the US Supreme Court." *Journal of Law, Economics, and Organization* 23(2):276–302.
- Lee, Frances E. 2016. *Insecure Majorities: Congress and the Perpetual Campaign*. Chicago: University of Chicago Press.
- Lemon, Andrew Yuichi. 2005. "Reputational Costs in Political Agency Models." Yale University PhD dissertation. Available at <http://www.princeton.edu/~smorris/past%20PhD%20Students> (Accessed 11 Oct 2018).
- Malhotra, Neil and Alexander G Kuo. 2008. "Attributing Blame: The Public's Response to Hurricane Katrina." *Journal of Politics* 70(1):120–135.
- Maskin, Eric and Jean Tirole. 2004. "The Politician and the Judge: Accountability in Government." *American Economic Review* 94(4):1034–1054.
- Patty, John W. 2016. "Signaling through Obstruction." *American Journal of Political Science* 60(1):175–189.

- Pear, Robert, Maggie Haberman and Reed Abelson. 2017. "Trump to Scrap Critical Health Care Subsidies, Hitting Obamacare Again." *The New York Times* Oct. 12th.
- Shotts, Kenneth and Scott Ashworth. 2011. "Challengers, Democratic Contestation, and Electoral Accountability." Stanford University working paper. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1901510](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1901510).
- Ting, Michael M. 2009. "Organizational Capacity." *The Journal of Law, Economics, & Organization* 27(2):245–271.
- Turner, Ian R. 2017. "Working Smart and Hard? Agency Effort, Judicial Review, and Policy Precision." *Journal of Theoretical Politics* 29(1):69–96.
- Warren, Patrick L. 2012. "Independent Auditors, Bias, and Political Agency." *Journal of Public Economics* 96(1-2):78–88.

# Appendix

**Notation:** For simplicity we henceforth write the voter's net policy benefit for the incumbent policy  $U(x_V; x_I, x_C)$  as  $U_V \geq 0$  which is assumed to be positive, and write the agent's net benefit for the incumbent policy  $U(x_S; x_I, x_C)$  as  $-U_S$ , where  $U_S \geq 0$  denotes the agent's net utility for the challenger policy. We also write  $\pi_0^0$  as just  $\pi_0$ ,  $\pi_1^1$  as  $\pi_H$ , and  $\pi_1^0$  as  $\pi_L$ . Finally, we suppress the explicit dependence of  $\bar{\Delta}_{\lambda_I}(\cdot)$  and  $\bar{\theta}_C(\cdot)$  on other quantities.

It is first helpful to show the property that  $\bar{\Delta}_H > \frac{q_H}{q_L} \bar{\Delta}_L$ , which furthermore has the implication that  $\bar{\Delta}_H \leq q_H \rightarrow \bar{\Delta}_L < q_L$ . This eliminates much of the parameter space and several potential types of equilibria. To see this, observe that the desired property is equivalent to

$$q_H \gamma + \delta \bar{\Delta}_L \frac{q_H}{q_L} \cdot (\gamma (q_H - q_L) (1 - \theta_C) - U_S) \geq 0$$

or

$$\frac{U_S - \gamma (q_H - q_L) (1 - \theta_C)}{U_S + \gamma \theta_C (q_H - q_L)} \leq 1$$

which clearly always holds.

## A Preliminary Analysis

Equilibrium values of  $e_L$  and  $e_H$  in conjunction with the incumbent's initial popularity imply different possible restrictions on the retention probabilities  $\pi_0$ ,  $\pi_L$ , and  $\pi_H$ . These in turn imply different feasible pairs of  $(\Delta_L, \Delta_H)$ . Anticipating these restrictions, we first examine several relevant feasible sets of  $(\pi_0, \pi_L, \pi_H)$  and their implications for  $(\Delta_L, \Delta_H)$ . Specifically, for each type of triple  $(\pi_0, \pi_L, \pi_H)$  we characterize feasible  $\Delta_L$  and then the feasible values of  $\Delta_H$  given  $\Delta_L$ . We then subsequently use this characterization in the equilibrium characterization.

In the subsequent case-by-base breakdown, (S) refers to “single mixing” (the voter mixes after one path of play) while (D) refers to “double-mixing” (the voter mixes after two paths of play).

**Case S.1** ( $\pi_0 \in (0, 1)$ ,  $\pi_L = \pi_H = 1$ ), We have

$$\Delta_{\lambda_I} = 1 - \pi_0$$

Therefore feasible values of  $\Delta_L$  are all  $\Delta_L \in [0, 1]$  and  $\Delta_H = \Delta_L$

**Case S.2** ( $\pi_0 = 0$ ,  $\pi_L = 0$ ,  $\pi_H \in (0, 1)$ ). We have

$$\Delta_{\lambda_I} = q_{\lambda_I} \pi_H$$

and it straightforward to show that  $\Delta_L \in [0, q_L]$  and  $\Delta_H = \frac{q_H}{q_L} \Delta_L$ .

**Case S.3** ( $\pi_0 = 0$ ,  $\pi_L \in (0, 1)$ ,  $\pi_H = 1$ ). We have

$$\Delta_{\lambda_I} = q_{\lambda_I} + (1 - q_{\lambda_I}) \pi_L$$

and it is straightforward to show that  $\Delta_L \in [q_L, 1]$  and  $\Delta_H = q_H + \left(\frac{1 - q_H}{1 - q_L}\right) (\Delta_L - q_L)$  which is clearly  $< \frac{q_H}{q_L} \Delta_L$ .

**Case S.4** ( $\pi_0 \in [0, 1]$ ,  $\pi_L = 0$ ,  $\pi_H = 1$ ). We have

$$\Delta_{\lambda_I} = q_{\lambda_I} - \pi_0$$

so it is straightforward that  $\Delta_L \in [0, q_L]$  and  $\Delta_H = \Delta_L + (q_H - q_L)$ .

**Case D.1** ( $\pi_0 \in [0, 1]$ ,  $\pi_L = 0$ ,  $\pi_H \in [0, 1]$ ). We have

$$\Delta_{\lambda_I} = -\pi_0 + q_{\lambda_I} \pi_H.$$

so it is straightforward that  $\Delta_L \in [0, q_L]$ . The potential values of  $\Delta_H$  then fall in an interval that we will characterize. The minimum possible value of  $\Delta_H$  occurs when  $\pi_0 = 0$  which is case **S.2** and so  $\Delta_H = \frac{q_H}{q_L} \Delta_L$ . The maximum possible value of  $\Delta_H$  occurs when when  $\pi_H = 1$ , which is **case S.4** and so the maximum value is  $\Delta_H = \Delta_L + (q_H - q_L)$ .

Summarizing, in Case D.1 we have  $\Delta_L \in [0, q_L]$  and  $\Delta_H \in \left[ \frac{q_H}{q_L} \Delta_L, \Delta_L + (q_H - q_L) \right]$

**Case D.2** ( $\pi_0 \in [0, 1]$ ,  $\pi_L \in [0, 1]$ ,  $\pi_H = 1$ ). We have

$$\Delta_{\lambda_I} = q_{\lambda_I} + (1 - q_{\lambda_I}) \pi_L - \pi_0$$

so it is straightforward that we may have any  $\Delta_L \in [0, 1]$ . The minimum possible value of  $\Delta_H$  occurs when  $\pi_L = 1$  which implies  $\Delta_H = \Delta_L$ . The maximum possible value of  $\Delta_H$  corresponds to the minimum possible value of  $\pi_L$ , which in turn depends on  $\Delta_L$ . If  $\Delta_L \in [0, q_L]$  then the minimum possible value of  $\pi_L$  is 0 and we are in case **S.4**, so  $\Delta_H = \Delta_L + (q_H - q_L)$ . If  $\Delta_L \in [q_L, 1]$  then the minimum possible value of  $\pi_L$  must be  $> 0$ ; the smallest feasible value corresponds with when  $\pi_0 = 0$ , so we are in case **S.3** and  $\Delta_H = q_H + \left( \frac{1 - q_H}{1 - q_L} \right) (\Delta_L - q_L)$ .

Summarizing, in case D.2 we have we have  $\Delta_L \in [0, 1]$  and

- if  $\Delta_L \in [0, q_L]$  then  $\Delta_H = [\Delta_L, \Delta_L + (q_H - q_L)]$
- if  $\Delta_L \in [q_L, 1]$  then  $\Delta_H = \left[ \Delta_L, q_H + \left( \frac{1 - q_H}{1 - q_L} \right) (\Delta_L - q_L) \right]$

## B Equilibrium Characterization

This section proceeds by enumerating all the types of equilibria and deriving existence conditions for each. After this analysis the equilibria are summarized as a function of the primitive parameters.

### B.1 Pooling on Effort Equilibria

We consider when pooling on effort is an equilibrium that satisfies D1 (Cho and Kreps 1987). Observe that when the voter observes sabotage, the only information he receives is from sabotage itself (since failure is assured). Consequently, when the saboteur is believed to be pooling on effort, any off-equilibrium path belief about the incumbent's type following sabotage  $\bar{\theta}_I^{0,0}(\cdot) \in [0, 1]$  satisfies sequential consistency (Kreps and Wilson 1982). Since the voter's reelection threshold  $\bar{\theta}_C(\cdot) \in (0, 1)$ , the voter's set of mixed best responses to consistent beliefs off the equilibrium path is *any* reelection probability  $\pi_0 \in [0, 1]$ . D1 thus requires the voter to assign probability weight 1 when one type of incumbent invites deviation for a strictly larger set of  $\pi_0 \in [0, 1]$ .

We now analyze the four popularity conditions.



**A very unpopular policy** ( $\bar{\theta}_C \in \left[ \frac{\theta_I q_H}{\theta_I q_H + (1-\theta_I) q_L}, 1 \right]$ ) We have  $\pi_H^* = \pi_L^* = 0$  and  $\Delta_L, \Delta_H \leq 0$ , so it is indeed an equilibrium to pool on effort regardless of the voters off-path best response ( $\pi_0^* \in [0, 1]$ ).

**A somewhat (un)popular policy** ( $\bar{\theta}_C \in \left[ \frac{\theta_I(1-q_H)}{\theta_I(1-q_H) + (1-\theta_I)(1-q_L)}, \frac{\theta_I q_H}{\theta_I q_H + (1-\theta_I) q_L} \right]$ ) Then  $\pi_H = 1 > \pi_L = 0$  and potential off path behavior is  $\pi_0 \in [0, 1]$ . Now we ask what different values of  $\pi_0$  imply for  $\Delta_L$  and  $\Delta_H$ —using case **S.3** the potential values of  $(\Delta_L, \Delta_H)$  are  $\Delta_L \in [0, q_L]$  and  $\Delta_H \in \Delta_L + (q_H - q_L)$ .

If  $\bar{\Delta}_L \geq q_L$  **then this is an equilibrium**; we know that this implies  $\bar{\Delta}_H \geq q_H$  and so no off path beliefs can invite deviation;  $\pi_0^*$  may be anything.

If  $\bar{\Delta}_L < q_L$ , then **this is an equilibrium i.f.f.**  $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$ . In this case, the set of  $\pi_0$  that invite deviation from a high type strictly contains the set that invite deviation from a low type, sabotage will be perceived as perfect good news (applying D1) and cause retention for sure so  $\pi_0^* = 1$ , and will therefore be undesirable.

Finally, if  $\bar{\Delta}_L < q_L$  but  $\bar{\Delta}_H > \bar{\Delta}_L + (q_H - q_L)$ , then again applying D1 sabotage will be perceived as bad news or  $\pi_0^* = 0$ , implying  $(\Delta_L = q_L, \Delta_H = q_H)$ , the bureaucrat will want to deviate to sabotaging both types, and this is not an equilibrium.

Summarizing, for a somewhat unpopular or somewhat popular policy, pooling on effort is an equilibrium i.f.f.

- $\bar{\Delta}_L \geq q_L$  or  $\bar{\Delta}_L < q_L$  and  $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$

Equilibrium retention probabilities are  $\pi_H^* = 1$ ,  $\pi_L^* = 0$ , and  $\pi_0^* = 1$ .

**A very popular policy** ( $\bar{\theta}_C \in \left[ 0, \frac{\theta_I(1-q_H)}{\theta_I(1-q_H) + (1-\theta_I)(1-q_L)} \right]$ ) No news and failure leads to retention ( $\pi_L = \pi_H = 1$ ). Then  $\Delta_{\lambda_I} = 1 - \pi_0$  so  $\Delta_L \in [0, 1]$  and  $\Delta_H = \Delta_L$ , the set of  $\pi_0$  that invite deviation from a bad type is always strictly larger than the set inviting deviation from a good type, sabotage should be perceived as *bad news* and cause the policy to be tossed for sure, so  $\pi_0^* = 0$ ,  $\Delta_L = \Delta_H = 1$ , sabotage will be desirable for both types and this is **not an equilibrium**.

**Summary** Pooling on effort is an equilibrium that satisfies D1 i.f.f.

- The policy is very unpopular, so  $\pi_H^* = \pi_L^* = 0$  and any  $\pi_0^*$
- The policy is somewhat unpopular or somewhat popular (so  $\pi_H^* = 1 > \pi_L^* = 0$ ), and either  $\bar{\Delta}_L \geq q_L$  (with any  $\pi_0^*$ ) or  $\bar{\Delta}_L < q_L$  and  $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$  (with  $\pi_0^* = 1$ )

## B.2 Pooling on Sabotage Equilibria

We consider when pooling on sabotage is an equilibrium that satisfies a modification of D1 (Cho and Kreps 1987). Specifically, when considering which incumbent type is more likely to invite deviation by the saboteur, we restrict attention to the set of off-equilibrium path mixed strategies by the voter that are best responses to sequentially consistent off-path beliefs (Kreps and Wilson 1982). Unlike the standard signalling game, sequential consistency imposes some constraints on the voter's off-equilibrium-path beliefs because nature sends an additional signal (success or failure) to the voter following the saboteur's move.

It is easily verified that when the saboteur is believed to be pooling on sabotage, any off equilibrium path belief about the incumbent's type  $\tilde{\theta}_I^1(\cdot) \in [0, 1]$  prior to the observation of success and failure satisfy sequential consistency. However, these beliefs will then be updated following success and failure using Bayes rule and the knowledge that high type incumbents succeed with probability  $q_H$  while low types succeed with probability  $q_L$ . The set of sequentially consistent beliefs following success and failure are thus

$$\tilde{\theta}_I^{1,1} = \frac{\tilde{\theta}_I^1 q_H}{\tilde{\theta}_I^1 q_H + (1 - \tilde{\theta}_I^1) q_L} \quad \text{and} \quad \tilde{\theta}_I^{1,0} = \frac{\tilde{\theta}_I^1 (1 - q_H)}{\tilde{\theta}_I^1 (1 - q_H) + (1 - \tilde{\theta}_I^1) (1 - q_L)}$$

for any value of  $\tilde{\theta}_I^1 \in [0, 1]$ . It is straightforward to verify that both  $\tilde{\theta}_I^{1,1}$  and  $\tilde{\theta}_I^{1,0}$  may each take any value  $\in [0, 1]$ , but  $\tilde{\theta}_I^{1,1} = \tilde{\theta}_I^{1,0}$  if and only if  $\tilde{\theta}_I^1 = 1$  or  $\tilde{\theta}_I^1 = 0$ ; otherwise  $\tilde{\theta}_I^{1,1} > \tilde{\theta}_I^{1,0}$ . Consequently, when the saboteur is believed to be pooling on sabotage, the voter's off-equilibrium-path set of mixed best responses to consistent beliefs following effort and success or failure are (i)  $\pi_L = 0$  and  $\pi_H \in [0, 1)$ , or (ii)  $\pi_L \in (0, 1]$  and  $\pi_H = 1$ .

We now analyze the four popularity conditions.

**An unpopular policy** ( $\bar{\theta}_C \leq \theta_P$ ) We argue pooling on sabotage is always an equilibrium. If the policy is unpopular then  $\pi_0^* = 0$ . Using that off-path actions are (i)  $\pi_L = 0$  and  $\pi_H \in [0, 1)$ , or (ii)  $\pi_L \in (0, 1]$  and  $\pi_H = 1$  straightforwardly yields the contour of impact probabilities  $\Delta_L \in [0, q_L]$  and  $\Delta_H = \frac{q_H}{q_L} \Delta_L$ , and  $\Delta_L \in [q_L, 1]$  and  $\Delta_H = q_H + \left(\frac{1-q_H}{1-q_L}\right) (\Delta_L - q_L)$  which is  $< \frac{q_H}{q_L} \Delta_L$ . Since we know  $\bar{\Delta}_H > \frac{q_H}{q_L} \bar{\Delta}_L$ , this implies the set of best responses inviting deviation from a high type is strictly larger than the set inviting deviation for a low type, implying effort should be interpreted as good news ( $\pi_H^* = \pi_L^* = 1$ ) and cause the policy to be retained for sure, and is therefore an undesirable deviation, so this is an equilibrium.

**A popular policy** ( $\bar{\theta}_C \geq \theta_P$ ) Then  $\pi_0 = 1$  and pooling on sabotage is not an equilibrium, since sabotage gets the policy retained for sure and also destroys valence.

## C (Partially) Separating Equilibria

We begin by ruling out certain types of strategy profiles.

First, we argue that  $(e_L > 0, e_H = 0)$  cannot be an equilibrium (including both  $e_L \in (0, 1)$  and  $e_L = 1$ , ruling out one type of separating equilibrium). Observe that effort is perfect bad news and causes policy to be tossed for sure ( $\pi_L = \pi_H = 0$ ), so it will be strictly desirable to exert effort for both types, contradicting  $e_H = 0$ .

We next argue that  $(e_L = 1, e_H < 1)$  cannot be an equilibrium. Observe that sabotage is perfect good news and causes the policy to be retained for sure ( $\pi_0 = 1$ ), so effort will weakly decrease the chance policy is retained, so again it will be strictly desirable on both types, contradicting  $e_H < 1$ .

Last we argue that  $(e_L = 0, e_H = 1)$  cannot be an equilibrium; combined with the above this rules out all separating equilibria. If so then effort perfectly reveals the incumbent is good while sabotage perfectly reveals the incumbent is bad; then  $\pi_H = \pi_L = 1$  and  $\pi_0 = 0$ , but then the bureaucrat will strictly prefer to sabotage a good incumbent under our assumptions, contradicting  $e_H = 1$ .

The remaining possible equilibrium efforts are four types of partially separating equilibria:

- (P1)  $e_L = 0, e_H \in (0, 1)$ : effort is “perfect good news,” sabotage is “imperfect bad news”
- (P2)  $e_L \in (0, 1), e_H = 1$ : effort is “perfect bad news,” sabotage is “noisy good news”
- (P3)  $0 < e_L < e_H < 1$ : effort is “noisy good news,” sabotage is “noisy bad news”
- (P4)  $0 < e_H < e_L < 1$ : effort is “noisy bad news,” sabotage is “noisy good news”

We consider each and derive conditions under which it is an equilibrium satisfying D1.

**(P1)**  $e_L = 0, e_H > 0$  Clearly  $\pi_H = \pi_L = 1$ . We first argue that for this to be an equilibrium requires the incumbent be popular or  $\theta_I \geq \bar{\theta}_C$ . If they are unpopular then  $\pi_0 = 0$  and  $\Delta_L = \Delta_H = 1$  and the agent will strictly prefer to sabotage a good policy, contradicting  $e_H > 0$ .

So suppose the incumbent is popular; we argue that it is always possible to derive an equilibrium of this form, and derive it. First, it is always possible to select  $e_H$  to generate principal indifference after sabotage generating  $\pi_0 \in [0, 1]$ , yielding case **S.1** from the preliminary analysis. This requires that

$$\bar{\theta}_C = \frac{\theta_I(1 - e_H)}{\theta_I(1 - e_H) + (1 - \theta_I)} \rightarrow e_H^* = \frac{\theta_I - \bar{\theta}_C}{(1 - \bar{\theta}_C)\theta_I}$$

Next, in S.1 we have  $\Delta_H = 1 - \pi_0$ , so to generate saboteur indifference with a high-ability incumbent requires

$$\Delta_H = \bar{\Delta}_H \iff \pi_0 = 1 - \bar{\Delta}_H$$

Finally, we have  $\Delta_L = \Delta_H = \bar{\Delta}_H > \bar{\Delta}_L$ , so the saboteur strictly prefers to sabotage a low-ability incumbent, supporting  $e_L = 0$ .

**(P2)**  $e_L \in (0, 1), e_H = 1$  We have  $\pi_0 = 0$ . We first argue this cannot be an equilibrium if the incumbent is very popular. If so, then  $\pi_H = \pi_L = 1$  (since effort is noisy good news), and the saboteur will strictly prefer to sabotage a high-ability incumbent, contradicting  $e_H = 1$ .

Next suppose that the incumbent is somewhat (un)popular, implying that  $\pi_H = 1$ . We argue an equilibrium of this form exists in which  $\pi_L \in (0, 1)$  i.f.f.  $\bar{\Delta}_L \in [q_L, 1]$ , and derive the equilibrium. First, it is always possible to select  $e_L$  to generate principal indifference

after effort and failure so that  $\pi_L \in (0, 1)$ , yielding case **S.3** from the preliminary analysis. This requires that

$$\bar{\theta}_C = \frac{\theta_I(1 - q_H)}{\theta_I(1 - q_H) + (1 - \theta_I)e_L(1 - q_L)} \rightarrow e_L^* = \frac{\theta_I(1 - q_H)}{(1 - \theta_I)(1 - q_L)} \bigg/ \frac{\bar{\theta}_C}{1 - \bar{\theta}_C}$$

Next, in S.3 we must have  $\Delta_L \in [q_L, 1]$  and  $\Delta_H = q_H + \left(\frac{1 - q_H}{1 - q_L}\right)(\Delta_L - q_L)$ , which is clearly  $< \frac{q_H}{q_L}\Delta_L$ . So  $\Delta_L = \bar{\Delta}_L \iff \bar{\Delta}_L \in [q_L, 1]$ , the desired necessary condition. To derive  $\pi_L$  observe that

$$\bar{\Delta}_L = q_L + (1 - q_L)\pi_L \iff \pi_L = \frac{\bar{\Delta}_L - q_L}{1 - q_L}$$

Finally,  $\Delta_H = q_H + \left(\frac{1 - q_H}{1 - q_L}\right)(\bar{\Delta}_L - q_L) < \frac{q_H}{q_L}\bar{\Delta}_L < \bar{\Delta}_H$ , so the saboteur strictly prefers to exert effort for a high ability incumbent, supporting  $e_H = 1$ .

Finally, suppose that the incumbent is very unpopular, so  $e_L$  may be chosen to generate principal indifference after both failure ( $\pi_L \in (0, 1)$  and  $\pi_H = 1$ ) or success ( $\pi_L = 0$  and  $\pi_H \in (0, 1)$ ). Using the analysis in the somewhat (un)popular case, an equilibrium of the former type exists i.f.f.  $\bar{\Delta}_L \in [q_L, 1]$ , and the equilibrium quantities are as previously derived. We now argue that an equilibrium of the latter type exists i.f.f.  $\bar{\Delta}_L \in [0, q_L]$ . We must select  $e_L$  to generate principal indifference after effort and success so that  $\pi_H \in (0, 1)$ , yielding case **S.2** from the preliminary analysis. This requires that

$$\bar{\theta}_C = \frac{\theta_I q_H}{\theta_I q_H + (1 - \theta_I)e_L q_L} \rightarrow e_L^* = \frac{\theta_I q_H}{(1 - \theta_I)q_L} \bigg/ \frac{\bar{\theta}_C}{1 - \bar{\theta}_C}$$

Next, in S.2 we must have that  $\Delta_L \in [0, q_L]$  and  $\Delta_H = \frac{q_H}{q_L}\Delta_L$ . So  $\Delta_L = \bar{\Delta}_L \iff \bar{\Delta}_L \in [0, q_L]$ , the desired necessary condition. To derive  $\pi_H$  observe that

$$\bar{\Delta}_L = q_L \pi_H \iff \pi_H = \frac{\bar{\Delta}_L}{q_L}$$

Finally,  $\Delta_H = \frac{q_H}{q_L}\bar{\Delta}_L < \bar{\Delta}_H$ , so the saboteur strictly prefers to exert effort for a high ability incumbent, supporting  $e_H = 1$ .

**Summary** There exists an equilibrium with  $e_H = 1$  and  $e_L \in (0, 1)$  i.f.f.

- The incumbent is very unpopular, somewhat unpopular, or somewhat popular and  $\bar{\Delta}_L \in [q_L, 1]$ . In the equilibrium

$$e_L^* = \frac{\theta_I(1 - q_H)}{(1 - \theta_I)(1 - q_L)} \bigg/ \frac{\bar{\theta}_C}{1 - \bar{\theta}_C}, \pi_0 = 0 < \pi_L = \frac{\bar{\Delta}_L - q_L}{1 - q_L} < \pi_H = 1$$

- The incumbent is very unpopular and  $\bar{\Delta}_L \in [0, q_L]$ . In the equilibrium

$$e_L^* = \frac{\theta_I q_H}{(1 - \theta_I)q_L} \bigg/ \frac{\bar{\theta}_C}{1 - \bar{\theta}_C}, \pi_0 = \pi_L = 0 < \pi_H = \frac{\bar{\Delta}_L}{q_L} < 1$$

**(P3)**  $0 < e_L < e_H < 1$  First observe that when both  $e_{\lambda_I} \in (0, 1) \forall \lambda_I$  we cannot have  $\pi_H \in (0, 1)$  and  $\pi_L \in (0, 1)$  since voter posterior beliefs after success are always strictly higher than posteriors after failure (unless effort is perfectly informative). Thus to generate saboteur mixing for both incumbent types requires  $\pi_0 \in (0, 1)$  and either  $0 = \pi_L < \pi_H < 1$  (case D.1) or  $0 < \pi_L < 1 = \pi_H$  (case D.2).

We first argue that for an equilibrium with  $0 < e_L < e_H < 1$  the following conditions are necessary and sufficient: (a) the incumbent is somewhat popular ( $\bar{\theta}_C \in \left[\theta_I, \frac{\theta_I q_H}{\theta_I q_H + (1 - \theta_I) q_L}\right]$ ), (b) reelection probabilities are as in case D.2 ( $0 < \pi_L < 1 = \pi_H$ ), (c)  $\bar{\Delta}_L \in [0, q_L]$ , and (d)  $\bar{\Delta}_H \in [\bar{\Delta}_L, \bar{\Delta}_L + (q_H - q_L)]$ .

If instead the incumbent were very popular then  $\pi_H = \pi_L = 1$ , a contradiction; if the incumbent were unpopular then  $\pi_0 = 0$ , also a contradiction. Finally, if the incumbent is somewhat popular then  $\pi_H = 1$ , so reelection probabilities must be as in case D.2.

Now if the incumbent is somewhat popular then it is always possible to select  $(e_L, e_H)$  to generate principal indifference after both sabotage and effort and failure. Equilibrium effort levels solve the following system of equations:

$$\begin{aligned} \frac{\theta_I e_H (1 - q_H)}{\theta_I e_H (1 - q_H) + (1 - \theta_I) e_L (1 - q_L)} &= \frac{1}{1 + \frac{(1 - \theta_I) e_L (1 - q_L)}{\theta_I e_H (1 - q_H)}} = \bar{\theta}_C \\ &= \frac{\theta_I (1 - e_H)}{\theta_I (1 - e_H) + (1 - \theta_I) (1 - e_L)} = \frac{1}{1 + \frac{(1 - \theta_I) (1 - e_L)}{\theta_I (1 - e_H)}} \end{aligned}$$

which yields

$$\frac{e_L}{e_H} = \left(\frac{1 - q_H}{1 - q_L}\right) \left(\frac{\theta_I}{1 - \theta_I} \Big/ \frac{\bar{\theta}_C}{1 - \bar{\theta}_C}\right) \quad \text{and} \quad \frac{1 - e_L}{1 - e_H} = \frac{\theta_I}{1 - \theta_I} \Big/ \frac{\bar{\theta}_C}{1 - \bar{\theta}_C}$$

Solving then yields

$$e_L^* = \left(\frac{1 - q_H}{q_H - q_L}\right) \frac{(\theta_I - \bar{\theta}_C)}{\bar{\theta}_C (1 - \theta_I)} \quad \text{and} \quad e_H^* = \left(\frac{1 - q_L}{q_H - q_L}\right) \frac{(\theta_I - \bar{\theta}_C)}{\theta_I (1 - \bar{\theta}_C)}.$$

Finally, for the saboteur to mix on both types of incumbents requires that  $\Delta_L = \bar{\Delta}_L$  and  $\Delta_H = \bar{\Delta}_H$ . We argue this implies  $\bar{\Delta}_L \in [0, q_L]$ , which in turn implies  $\bar{\Delta}_H \in [\bar{\Delta}_L, \bar{\Delta}_L + (q_H - q_L)]$  from the preliminary analysis of case D.2. If instead  $\bar{\Delta}_L \in [q_L, 1]$  then we must have  $\bar{\Delta}_H \in \left[\bar{\Delta}_L, q_H + \left(\frac{1 - q_H}{1 - q_L}\right) (\bar{\Delta}_L - q_L)\right]$  (again from the preliminary analysis), but  $\bar{\Delta}_H > \frac{q_H}{q_L} \bar{\Delta}_L > q_H + \left(\frac{1 - q_H}{1 - q_L}\right) (\bar{\Delta}_L - q_L)$ , a contradiction. Finally, in case D.2 the retention probabilities are defined by the system  $(\pi_L - \pi_0) + q_{\lambda_I} (1 - \pi_L) = \bar{\Delta}_{\lambda_I} \forall \lambda_I$  and we have

$$\pi_L = \frac{(q_H - \bar{\Delta}_H) - (q_L - \bar{\Delta}_L)}{q_H - q_L} \quad \text{and} \quad \pi_0 = \frac{(1 - q_L) (q_H - \bar{\Delta}_H) - (1 - q_H) (q_L - \bar{\Delta}_H)}{q_H - q_L}$$

**(P4)**  $0 < e_H < e_L < 1$  We first argue that: (a) the incumbent must be somewhat unpopular ( $\bar{\theta}_C \in \left[\frac{\theta_P (1 - q_H)}{\theta_P (1 - q_H) + (1 - \theta_P) (1 - q_L)}, \theta_I\right]$ ), (b) reelection probabilities are as in case D.1 ( $\pi_0 \in (0, 1)$  and  $0 = \pi_L < \pi_H < 1$ ), (c)  $\bar{\Delta}_L \in [0, q_L]$ , and (d)  $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$ .

As in the analysis in (P3) we must have  $\pi_0 \in (0, 1)$  and either  $0 = \pi_L < \pi_H < 1$  (case D.1) or  $0 < \pi_L < 1 = \pi_H$  (case D.2). If the incumbent were very unpopular then we would have  $\pi_L = \pi_H = 0$ , a contradiction; if she were popular we would have  $\pi_0 = 1$ , also a contradiction;

she must therefore be somewhat unpopular, further implying  $0 = \pi_L < \pi_H < 1$  (case D.1).

Now if the incumbent is somewhat unpopular then it is always possible to select  $(e_L, e_H)$  to generate principal indifference after both sabotage and effort and failure. Equilibrium effort levels solve the following system of equations:

$$\begin{aligned} \frac{\theta_I e_H q_H}{\theta_I e_H q_H + (1 - \theta_I) e_L q_L} &= \frac{1}{1 + \frac{(1 - \theta_I) e_L q_L}{\theta_I e_H q_H}} = \bar{\theta}_C \\ &= \frac{\theta_I (1 - e_H)}{\theta_I (1 - e_H) + (1 - \theta_I) (1 - e_L)} = \frac{1}{1 + \frac{(1 - \theta_I) (1 - e_L)}{\theta_I (1 - e_H)}} \end{aligned}$$

which yields

$$\frac{e_L}{e_H} = \frac{q_H}{q_L} \cdot \left( \frac{\theta_I}{1 - \theta_I} \Big/ \frac{\bar{\theta}_C}{1 - \bar{\theta}_C} \right) \quad \text{and} \quad \frac{1 - e_L}{1 - e_H} = \frac{\theta_I}{1 - \theta_I} \Big/ \frac{\bar{\theta}_C}{1 - \bar{\theta}_C}$$

Solving yields the interior solution

$$e_L^* = \left( \frac{q_H}{q_H - q_L} \right) \frac{(\bar{\theta}_C - \theta_I)}{\bar{\theta}_C (1 - \theta_I)} \quad \text{and} \quad e_H^* = \left( \frac{q_L}{q_H - q_L} \right) \frac{(\bar{\theta}_C - \theta_I)}{\theta_I (1 - \bar{\theta}_C)}$$

Finally, for the saboteur to mix on both types of incumbents requires that  $\Delta_L = \bar{\Delta}_L$  and  $\Delta_H = \bar{\Delta}_H$ . From the preliminary analysis of case D.1 this immediately implies  $\bar{\Delta}_L \in [0, q_L]$  and  $\bar{\Delta}_H \in [\bar{\Delta}_L, \bar{\Delta}_L + (q_H - q_L)]$ . The retention probabilities are defined by the system  $-\pi_0 + q_{\lambda_I} \pi_H = \bar{\Delta}_{\lambda_I} \forall \lambda_I$  which yields

$$\pi_H^* = \frac{\bar{\Delta}_H - \bar{\Delta}_L}{q_H - q_L} \quad \text{and} \quad \pi_0^* = \frac{q_L \bar{\Delta}_H - q_H \bar{\Delta}_L}{q_H - q_L}$$

## D Additional Proofs

We now provide additional proofs that support stated results in the main text.

**Sufficient condition for  $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$**

We prove that the inequality in the equilibrium statements for a somewhat (un)popular is a sufficient condition for both  $\bar{\Delta}_L \leq q_L$  and  $\bar{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$ . These latter properties substantially simplify the equilibrium characterization by eliminating many possibilities.

From the definitions we have that

$$q_{\lambda_I} = \delta \bar{\Delta}_{\lambda_p} \left( \frac{U_S}{\gamma_S} - (\mathbf{1}_{\lambda_I=H} - \theta_C) (q_H - q_L) \right)$$

which is equivalent to

$$q_{\lambda_I} + \delta \bar{\Delta}_{\lambda_p} (q_H - q_L) = \delta \bar{\Delta}_{\lambda_p} \left( \frac{U_S}{\gamma_S} + \theta_C (q_H - q_L) \right)$$

Also observe that that  $\bar{\Delta}_{\lambda_p} \leq \Delta_{\lambda_p} \iff$

$$\frac{U_S}{\gamma_S} \geq (\mathbf{1}_{\lambda_I=H} - \theta_C) (q_H - q_L) + \frac{1}{\delta} \frac{q_{\lambda_I}}{\Delta_{\lambda_p}}$$

Now define  $\hat{\Delta}_H$  to be the quantity satisfying

$$q_H + \delta q_H (q_H - q_L) = \delta \hat{\Delta}_H \left( \frac{U_S}{\gamma_S} + \theta_C (q_H - q_L) \right)$$

or

$$q_H (1 + \delta (q_H - q_L)) = \delta \hat{\Delta}_H \left( \frac{U_S}{\gamma_S} + \theta_C (q_H - q_L) \right)$$

From the definitions, any value of  $\hat{\Delta}_H$  corresponding to a value of  $\bar{\Delta}_H < q_H$  must satisfy  $\bar{\Delta}_H < \hat{\Delta}_H$ . It is also straightforward to see that

$$\frac{\hat{\Delta}_H}{\bar{\Delta}_L} = \frac{q_H (1 + \delta (q_H - q_L))}{q_L} \iff \hat{\Delta}_H = \frac{q_H (1 + \delta (q_H - q_L))}{q_L} \bar{\Delta}_L$$

We now consider when we have  $\hat{\Delta}_H \leq \bar{\Delta}_L + (q_H - q_L)$ ; this requires

$$\frac{q_H (1 + \delta (q_H - q_L))}{q_L} \bar{\Delta}_L \leq \bar{\Delta}_L + (q_H - q_L) \iff \bar{\Delta}_L \leq \frac{q_L}{1 + q_H \delta}$$

(which is stronger than  $\bar{\Delta}_L \leq q_L$ ). From the definition this condition is equivalent to:

$$\frac{U_S}{\gamma_S} \geq \frac{1}{\delta} + (1 - \theta_C) (q_H - q_L) + q_L$$

Further, it is also easily verified that  $\bar{\Delta}_H \leq q_H \iff$

$$\frac{U_S}{\gamma_S} \geq \frac{1}{\delta} + (1 - \theta_C) (q_H - q_L)$$

which is a weaker condition, so when the stated condition holds we have  $\bar{\Delta}_H < \hat{\Delta}_H < \bar{\Delta}_L + (q_H - q_L)$  and this is sufficient for the desired properties. Finally, if we would like the condition to hold for *all* values of  $\theta_C$  then we require  $\frac{U_S}{\gamma_S} \geq \frac{1}{\delta} + q_H$ .

### Proof of Proposition 1

Sequential equilibrium (Kreps and Wilson 1982) straightforwardly implies that both on and off the equilibrium path, the voter's beliefs will be computed using Bayes' rule using nature's probabilities of success and failure and ignoring the behavior of the saboteur. Optimal behavior is thus straightforwardly described by the popularity conditions.

To see the incumbent strategy, it is straightforward that the saboteur will never sabotage when the incumbent is very (un)popular (since doing so would have no effect on the probability of retention) or when the saboteur is somewhat popular (since sabotage would be counterproductive and ensure retention).

If the incumbent is somewhat unpopular, the net benefit of exerting effort simply the expected value of the net benefit for each incumbent type:

$$(1 - \theta_I) (q_L \gamma_S + \delta q_L (V(\mathbf{0}, \theta_C; \gamma_S, q) + U(x_S; x_I, x_C))) \\ + \theta_I (q_H \gamma_S + \delta q_H (V(\mathbf{1}, \theta_C; \gamma_S, q) + U(x_S; x_I, x_C)))$$

and the saboteur will sabotage i.f.f. this is  $\leq 0$ .

This expression may be rewritten as

$$((1 - \theta_I) q_L + \theta_I q_H) \left( \frac{1}{\delta} - \theta_C (q_H - q_L) + \frac{U(x_S; x_I, x_C)}{\gamma_S} \right) + \theta_I q_H (q_H - q_L) \leq 0$$

which in turn is easily rearranged to the expression in the proposition.

### Proof of Propositions 2-5

By the equilibrium characterization and the assumption that  $-\frac{U(x_S; x_I, x_C)}{\gamma_S} \geq \frac{1}{\delta} + q_H$  there are three equilibria satisfying D1: (a) pooling on sabotage, (b) pooling on effort, and (c) the partially separating equilibrium (P4) with  $0 < e_H < e_L < 1$ .

*Saboteur*

We first show that the saboteur prefers pooling on sabotage to (P4) to pooling on effort.

To see that the saboteur strictly prefers pooling on sabotage to (P4), observe that deviating from her P4 strategy profile to pooling on sabotage yields her equilibrium utility due to the equilibrium indifference conditions; however, this involves the incumbent retained with strictly positive probability, and is therefore strictly worse than the equilibrium with pooling on sabotage in which the incumbent is replaced for sure.

To see that the saboteur strictly prefers (P4) to pooling on effort, observe that deviating from her (P4) strategy to pooling on effort yields her equilibrium utility, but the incumbent is retained after success with probability  $\pi_H^{P4} < 1$ ; this is thus strictly better than the equilibrium with pooling on effort in which an incumbent who succeeds is retained for sure.

*Voter*

We now show that the voter prefers pooling on effort to P4 to pooling on sabotage.

To see that the voter strictly prefers pooling on effort to (P4), we make a sequence of changes altering the strategy profile in (P4) to that in the pooling on effort equilibrium that each weakly increase her utility. First, changing from  $(\pi_0^{P4} \in (0, 1), \pi_L^{P4} = 0, \pi_H^{P4} \in (0, 1); e_L^{P4}, e_H^{P4})$  to  $(\pi_0 = \pi_L = \pi_H = 0; e_L^{P4}, e_H^{P4})$  does not change the voter's utility due to the (P4) indifference conditions. Next changing to  $(\pi_0 = \pi_L = \pi_H = 0; e_L = e_H = 1)$  strictly increases the voters's utility since first period quality increases with no change in selection. Finally, changing to  $(0 = \pi_L = \pi_0 < \pi_H = 1; e_L = e_H = 1)$  strictly increases the voter's utility since retention is strictly optimal after success when effort is uninformative.

To see that the voter strictly prefers (P4) to pooling on sabotage, observe that a deviation in (P4) to  $\pi_0 = \pi_L = \pi_H = 0$  (always replace) does not change her utility, which involves strictly positive effort levels; this is thus strictly better for the voter than the pooling on sabotage equilibrium which also involves always replacing, but with no effort.

**Proof of Proposition 6**

From the equilibrium characterization, we have that:

$$\frac{e_L}{e_H} = \frac{q_H}{q_L} \cdot \left( \frac{\theta_I}{1 - \theta_I} \bigg/ \frac{\bar{\theta}_C}{1 - \bar{\theta}_C} \right) \quad \text{and} \quad \frac{1 - e_H}{1 - e_L} = \frac{\bar{\theta}_C}{1 - \bar{\theta}_C} \bigg/ \frac{\theta_I}{1 - \theta_I}$$

The ratio  $\frac{e_L}{e_H} \geq 1$  reflects the extent to which effort is “bad news” while the ratio  $\frac{1 - e_H}{1 - e_L} \geq 1$  reflects the extent to which sabotage is “good news.” Now let  $R(\bar{\theta}_C, \theta_I) = \frac{\bar{\theta}_C}{1 - \bar{\theta}_C} \bigg/ \frac{\theta_I}{1 - \theta_I}$ ; it is easily verified that this increases from 1 to  $\frac{q_H}{q_L}$  as  $\bar{\theta}_C$  increases from  $\theta_I$  to  $\bar{\theta}_I^1$ . Rewriting we have that:

$$\frac{e_L}{e_H} = \frac{q_H/q_L}{R(\bar{\theta}_C, \theta_I)} \quad \text{and} \quad \frac{1 - e_H}{1 - e_L} = R(\bar{\theta}_C, \theta_I)$$

First observe by multiplying the two equations that:

$$\frac{e_L}{1 - e_L} \bigg/ \frac{e_H}{1 - e_H} = \frac{q_H}{q_L}$$

This immediately yields that  $e_L$  and  $e_H$  must move strictly in the same direction as a function of  $R(\bar{\theta}_C, \theta_I)$ ; otherwise the LHS could not be constant.



Next observe that  $e_H = R(\bar{\theta}_C, \theta_I) \frac{q_L}{q_H} e_L$  and  $1 - e_H = R(\bar{\theta}_C, \theta_I) (1 - e_L)$  so summing the equations yields:

$$1 = R(\bar{\theta}_C, \theta_I) \left( 1 - \left( 1 - \frac{q_L}{q_H} \right) e_L \right)$$

Thus  $e_L$  (and from the preceding  $e_H$ ) are strictly increasing in  $R(\bar{\theta}_C, \theta_I)$ , which is in turn strictly decreasing in  $\theta_I$  and strictly increasing in  $\bar{\theta}_C$ , which in turn is strictly increasing in  $\theta_C$  and  $\gamma_V$  and strictly decreasing in  $U(x_V; \cdot)$ .

### Proof of Proposition 7

Recall that

$$\bar{\Delta}_{\lambda_i} = \frac{q_{\lambda_I}}{\delta(B - 1_{\lambda_I=H}(q_H - q_L))}$$

where  $B = \delta \left( \frac{-U(x_S; x_I, x_C)}{\gamma_S} + \theta_C (q_H - q_L) \right) > q_H - q_L$  by assumption (so  $\bar{\Delta}_H < 1$ ). Now from the equilibrium characterization we have that

$$\pi_H = \frac{\bar{\Delta}_H - \bar{\Delta}_L}{q_H - q_L} \quad \text{and} \quad \pi_0 = \frac{q_L \bar{\Delta}_H - q_H \bar{\Delta}_L}{q_H - q_L}$$

Substituting in the definitions and algebra yields that

$$\pi_H = \frac{1 + \frac{q_L}{B}}{\delta(B - (q_H - q_L))} \quad \text{and} \quad \pi_0 = \frac{q_L q_H}{\delta(B - (q_H - q_L)) B}.$$

Both quantities are straightforwardly decreasing in  $B$  and  $\delta$ . It is also easily verified that

$$\pi_H - \pi_0 = \frac{1 + \frac{q_L(1 - q_H)}{B}}{\delta(B - (q_H - q_L))}$$

Thus all three quantities are straightforwardly decreasing in  $B$  and  $\delta$ .

### Proof of Proposition 8

Follows immediately from the equilibrium characterization.

### Proof of Proposition 9

By the equilibrium characterization there are three equilibria satisfying D1: (1) pooling on effort, (2) pooling on sabotage, and (3) the partially separating equilibrium (P2) with  $e_H = 1$  and  $e_L \in (0, 1)$ ; the assumption also yields  $0 = \pi_0 = \pi_L < \pi_H < 1$ .

We now argue that pooling on effort is Pareto dominant. Pareto dominance of pooling on effort to pooling on sabotage is straightforward; both involve the incumbent being replaced with probability 1, and holding retention decisions fixed both players prefer higher effort to lower effort.

We next compare pooling on effort to (P2). With pooling on effort, we have  $\pi_L = \pi_H = 0$  and the incumbent is always replaced. In (P2), we have equilibrium  $(\pi_0^*, \pi_L^*, \pi_H^*)$  and  $(e_L^*, e_H^*)$ . To see that the saboteur strictly prefers the equilibrium with pooling on effort, observe that the retention probabilities yield indifference over effort on a low quality incumbent, so the saboteur gets the same utility by deviating to pooling on effort ( $e_L = 1, e_H = 1$ ) with  $(\pi_0^*, \pi_L^*, \pi_H^*)$ , which involves retention with strictly positive probability and is therefore strictly worse.

To see that the voter strictly prefers the equilibrium with pooling on effort, observe that the voter still gets her equilibrium utility by deviating to always replace given the

bureaucrat's equilibrium effort levels, which in turn is worse than always replacing with maximum effort by the bureaucrat.

### Proof of Proposition 10

By the equilibrium characterization and the assumption that  $-\frac{U(x_S; x_I, x_C)}{\gamma_S} \geq \frac{1}{\delta} + q_H$  there are three equilibria satisfying D1: (1) the partially separating equilibrium (P1) with  $e_L = 0$  and  $e_H \in (0, 1)$ ,  $0 < \pi_0 < 1 = \pi_L = \pi_H$ , (2) pooling on effort ( $0 = \pi_L < \pi_H = 1$ ), and (3) the partially separating equilibrium (P3) with  $0 < e_L < e_H < 1$  and  $\pi_0 \in (0, 1)$ ,  $\pi_L \in (0, 1)$ ,  $\pi_H = 1$ .

We first compare pooling on effort to (P3). For the saboteur, in (P3) a deviation to pooling on effort would still yield her equilibrium utility but with  $\pi_L^* > 0$ , so her equilibrium utility is strictly worse in (P3).

For the voter, we make a sequence of changes altering the strategy profile in (P3) to that in the pooling on effort that each weakly increase her equilibrium utility. First, changing from  $(\pi_0^{P3} \in (0, 1), \pi_L^{P3} \in (0, 1), \pi_H = 1; e_L^{P3}, e_H^{P3})$  to  $(\pi_0 = \pi_L = \pi_H = 1; e_L^{P3}, e_H^{P3})$  does not change the voter's utility due to the (P3) indifference conditions. Next changing to  $(\pi_0 = \pi_L = \pi_H = 1; e_L = e_H = 1)$  strictly increases the voters's utility since first period quality increases with no change in selection. Finally, changing to  $(0 = \pi_L < \pi_0 = \pi_H = 1; e_L = e_H = 1)$  strictly increases the voter's utility since replacement is strictly optimal after failure when effort is uninformative.

We next compare (P3) to (P1). For the saboteur, a deviation to the both equilibrium effort levels in (P1) would yield her (P3) equilibrium utility holding retention probabilities fixed. We next argue that the equilibrium retention probabilities in P1 are uniformly higher, implying that the saboteur is worse off in the (P1) equilibrium than in the (P3) equilibrium. Clearly retention probabilities are higher in (P1) after success and failure; we need only argue that the retention probability is also higher after sabotage. From the equilibrium characterizations we have that

$$\pi_0^{P1} = 1 - \bar{\Delta}_H \text{ and } \pi_0^{P3} = (q_H + (1 - q_H)\pi_L^{P3}) - \bar{\Delta}_H$$

which shows the desired property since  $q_H + (1 - q_H)\pi_L^{P3} < 1$ .

For the voter, a deviation to always retain in (P1) still yields her (P1) equilibrium utility, and a deviation to always retain in (P3) still yields her (P3) equilibrium utility. Thus, it suffices to show  $e_{\lambda_P}^{P3} > e_{\lambda_P}^{P1} \forall \lambda_P$ . We immediately have  $e_L^{P3} > 0 = e_L^{P3}$ . In addition, in both equilibria  $\pi_0 \in (0, 1)$  requires

$$\frac{\theta_I(1 - e_H)}{(1 - \theta_I)(1 - e_L)} = \frac{\bar{\theta}_C}{1 - \bar{\theta}_C},$$

but this immediately yields  $e_L^{P3} > e_L^{P1} \rightarrow e_H^{P3} > e_H^{P1}$ .