THE UNIVERSITY OF CHICAGO

ESSAYS ON THE THEORY OF COLLECTIVE PUNISHMENT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
DEPARTMENT OF POLITICAL SCIENCE
AND
THE FACULTY OF THE IRVING B. HARRIS
GRADUATE SCHOOL OF PUBLIC POLICY STUDIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN POLITICAL ECONOMY

BY
GIORGIO GIOVANNI FARACE

CHICAGO, ILLINOIS
AUGUST 2026

Abstract

This dissertation consists of three essays on the strategic logic of collective punishment.

The first essay studies why rulers knowingly punish the innocent. A ruler who can perfectly observe who obeys or disobeys designs a directed graph of collective responsibility, specifying who is punished for whose disobedience. Making people responsible for others trades *vertical* for *horizontal* deterrence: it reduces the protection generated by one's own obedience, but can induce people to police each other.

The second essay studies why rulers punish people for failure to report their disobedient peers. Now the ruler observes individual disobedience independently with some probability. When reporting is costly, punishing silence elicits information only under *two-sided ignorance*: the ruler must not know what people know, namely who disobeyed, and people must not know what the ruler already knows. Denunciation regimes are therefore optimal only at intermediate detection probabilities.

The third essay studies collective punishment when the ruler observes only how many people disobey. The obedience-maximizing rule is a threshold chosen to maximize individual pivotality. Such a threshold must be *self-modal*: from any person's perspective, it is the most likely number of other disobedient people under the equilibrium distribution that the threshold itself induces. As a result, collective punishment can be frequent even when individual disobedience is rare.

A mio papà.

The bearers who had lost their standards, as well as the centurions and previously decorated soldiers who had deserted the ranks, he ordered scourged and beheaded; *from the rest* every tenth man was selected by lot for execution.

Livy, *Ab Urbe Condita 2.59*, emphasis added

Table of contents

List of figures

Acknowledgments

I'm deeply grateful to my committee, whose example of mentorship I hope to carry with me forever. Scott Gehlbach's exceptional generosity and attention to both the details of a model and the larger questions behind it helped me see my scattered ideas as the beginning of a research agenda. Zhaotian Luo always welcomed unfinished models at his whiteboard and patiently helped me move forward. Scott Ashworth has been a model of clarity in thought and language; when revising an argument, I often ask how he would make it more precise. Konstantin Sonin's candor and ambition pushed me to ask more of my work. I owe them a great deal for their support on the job market.

This dissertation grew out of the extraordinary community of political economists in and around Chicago and its ruthless and generous seminar culture. I have learned so much from Wiola Dziuda and Ethan Bueno de Mesquita's way of doing theory and benefited greatly from their constant advice and feedback. My path to Chicago owes much to Andy Eggers, whose mentorship at Oxford brought me here and whose questions first set this dissertation in motion. Massimo Morelli was my first academic mentor; at Bocconi, he introduced me to game theory and the possibility of a life in research. My fellow Ph.D. students were a crucial part of this community and of my life in Chicago. I'm particularly thankful to Pepi Pandiloski and Ben Shaver for their friendship and advice.

My mom has always encouraged my intellectual pursuits and accepted with love the distance they required. My dad first got me interested in politics and taught me how to agree to disagree. Finally, I would have accomplished much less without Arshad, who made life around the Ph.D. happy and full. What a blessing to share a life with you!

1 General introduction

This dissertation studies the strategic logic of collective punishment, a sanction imposed on people for the actions of others. Throughout history, rulers have often punished collectively, sometimes because they cannot condition sanctions on individual behavior, and sometimes even when they can. Neighborhoods or villages are destroyed after acts of resistance; prison cells or blocks are locked down after individual infractions; whole classrooms are penalized for the conduct of a few. Despite the prevalence of these practices, the incentives they generate remain under-theorized. I address this gap with three formal models, each isolating a distinct mechanism through which collective punishment shapes incentives for obedience.

The three models share a common framework. A ruler faces a finite set of people $N := \{1, \ldots, n\}$ and publicly commits to a *punishment rule*, mapping what he observes into bounded sanctions. Each person i chooses whether to obey or not, where $a_i = 0$ denotes obedience and $a_i = 1$ disobedience. The ruler wants obedience, but disobedience may be tempting: people each have a private type measuring their direct gain from disobeying. The models differ in what the ruler observes, what he can choose, and whether people can police or report one another. Figure 1.1 summarizes the structure of the dissertation.

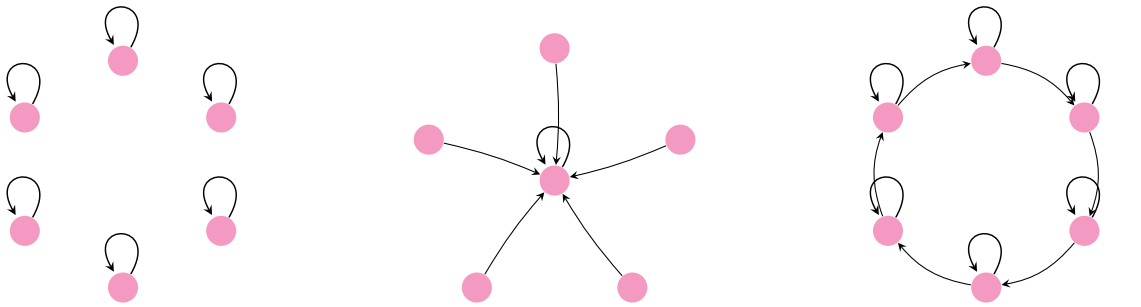
The first essay asks why rulers punish people for others' actions when individual conduct is observable. It develops a non-informational theory of collective punishment: rulers use collective responsibility to create incentives for people to sanction one another's disobedience. I model collective responsibility as a binary relation (or directed graph) R on the set of people: an edge iRj means that person i is punished when person j disobeys.

Essay	**Ruler sees**	**Ruler chooses**	**Peer action**	**Mechanism**
Responsibility	individual actions $(a_i)_{i \in N}$	responsibility relation $R \subseteq N \times N$	costly sanction	horizontal–vertical deterrence
Denunciations	random subset of disobedients + reports	denunciation regime, i.e., a partition of N	costly report	two-sided ignorance
Anonymity	disobedient count $\sum_{i \in N} a_i$	punishment rule $p \in [0,1]^{n+1}$	none	individual pivotality

Figure 1.1. Schematic overview of the dissertation.

The ruler faces a trade-off between *vertical deterrence*, generated by the direct threat of his punishment, and *horizontal deterrence*, generated by peer sanctions. Responsibility for others weakens vertical deterrence because obedience no longer guarantees safety from the ruler's sanction, but it can generate horizontal deterrence by inducing people to police others. A person may police someone for whom she is not directly responsible: it is enough that sanctioning that peer is not too costly and that the peer's obedience helps deter someone whose disobedience would expose her to punishment. Policing incentives follow these directed paths in the responsibility graph.

Figure 1.2 illustrates three responsibility relations which will be central to the analysis. When peer policing is too costly for the ruler to generate horizontal deterrence, individual responsibility is uniquely optimal. When horizontal deterrence can be generated, the ruler maximizes one person's obedience by making everyone else responsible for her, while making her responsible only for herself. Everyone has reason to police her, and her vertical deterrence remains maximal. However, maximal obedience is exclusive: the ruler cannot maximize two people's obedience at once. If the ruler must treat people symmetrically, and peer policing is sufficiently cheap, the optimal rule is maximally sparse: each person is responsible for herself and for exactly one other person. This rule gives everyone a reason to police everyone else while minimizing the loss in vertical deterrence.



(a) Individual responsibility (b) Star responsibility (c) Cycle responsibility

Figure 1.2. Three examples of collective responsibility as a directed graph. Arrows point from the person punished to the person whose disobedience triggers punishment. Panel (a) depicts individual responsibility: each person is responsible only for herself. Panel (b) depicts star responsibility: everyone is responsible for the person in the center. Panel (c) depicts cycle responsibility: each person is responsible for herself and one successor.

The second essay studies peer denunciations. It asks why rulers punish people not only for disobeying, but also for failing to report others' disobedience. The ruler independently observes each act of disobedience with a fixed probability, which measures his monitoring capacity. People perfectly observe one another's behavior and may report disobedience at a cost. The ruler chooses a denunciation regime: a partition of N into reporting groups. Within a group, a person is punished when she is caught by the ruler, when she is reported by someone, or when someone in her group was caught by the ruler but not reported. A degenerate case is the partition into singletons, where people are never expected to denounce peers; this is the analogue of individual responsibility in the first essay.

The mechanism is *two-sided ignorance*. People may know about violations that the ruler misses, but they do not know which violations the ruler has already detected. The mechanism benefits the ruler only if his monitoring capacity is intermediate. When it is too low, the ruler rarely observes violations independently, making reports hard to incentivize; when it is too high, direct monitoring already deters disobedience, so reporting duties mostly add costs that weaken obedience incentives. Strikingly, even as reporting costs vanish, optimal group size remains bounded above by the inverse of the ruler's monitoring capacity. Finally,

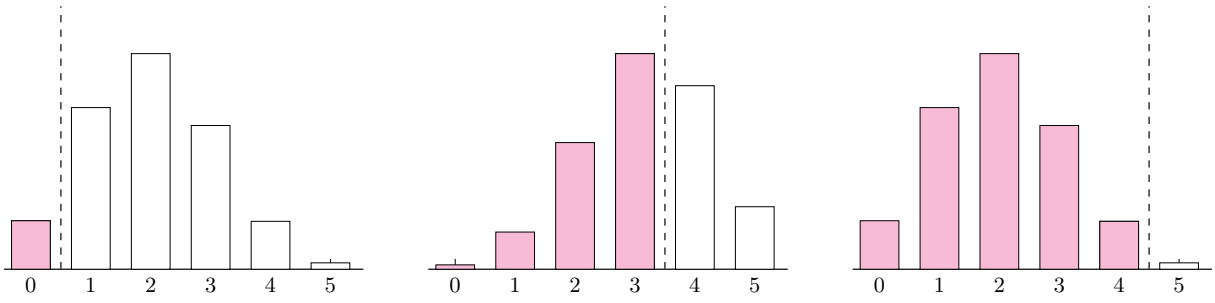
if people observe one another only imperfectly, the same two-sided ignorance logic can create incentives for baseless accusations.

The third essay removes peer actions altogether and studies collective punishment under anonymity. The ruler cannot condition punishment on individual identities: he observes only the number of people who disobey and must punish everyone equally. Here, a punishment rule maps the number of disobedients to a collective punishment in the unit interval; hence it is a function $p : \{0, \ldots, n\} \to [0, 1]$ (or equivalently a vector $p \in [0, 1]^{n+1}$). Incentives can be created only through pivotality: a person obeys because her obedience may keep the group from being punished.

I show that an optimal punishment rule is a threshold rule: the group is punished if and only if total disobedience is sufficiently high. For any given threshold, people respond to the incentives it creates, generating an equilibrium distribution of the number of disobedients. Figure 1.3 illustrates this logic. The optimal threshold is *self-modal*: from any person's perspective, the pivotal count is also the most likely number of other people to disobey under the equilibrium distribution induced by that same threshold. This result explains why frequent punishment is a structural feature of optimal deterrence under aggregate monitoring, not necessarily evidence of weak control or latent disloyalty. The ruler places the punishment boundary where a person's action is most likely to matter; that also means the group is often close to the boundary, so punishment occurs frequently.

The models also rest on several common assumptions. In each, people's gains from disobedience are private and independently distributed. The ruler commits to a punishment rule before people act and maximizes obedience, ignoring any cost of punishing. Fixing the ruler's institutional design, whenever there are multiple equilibria, I generally select the one that yields the highest expected obedience. These assumptions are made for both analytical convenience and theoretical parsimony.

From the people's perspective, the private-value and independence assumptions fix the source of strategic interdependence. A person's payoff from disobedience depends on her



(a) Threshold too low (b) Self-modal threshold (c) Threshold too high

Figure 1.3. Each panel shows a threshold rule and the equilibrium distribution relevant for one person's pivotality. Pink bars mark outcomes that lead to punishment. The self-modal threshold places the punishment boundary where one person's action is most likely to determine whether the group is punished.

own type, not directly on other people's types or actions. Other people's choices matter only through the ruler's punishment rule. Thus, whenever obedience decisions are strategically linked, the link comes from the ruler's institutional design.

From the ruler's perspective, even in a frictionless world in which he can fully commit and punishing is free, optimal behavior is often restrained. In the first essay, maximizing obedience can require exclusive or sparse responsibility relations. In the second, optimal reporting groups remain bounded even as reporting costs vanish. In the third, the optimal rule is self-modal, rather than a zero-tolerance rule under which any disobedience triggers punishment. This restraint is solely a consequence of the incentive structures that can be generated by collective punishment. Extreme forms of indiscriminate collective punishment therefore require explanations other than obedience maximization.

2 Repression by collective responsibility

> In the camps things are arranged so that the *zek* is kept up to the mark not by his bosses but by the others in his gang. Either everybody gets a bonus or else they all die together.
>
> Aleksandr Solzhenitsyn, *One Day in the Life of Ivan Denisovich*

2.1 Why punish the innocent?

It is generally accepted that people should be accountable only for their own actions.[1] Collective punishments violate this principle and are forbidden by the laws of war; yet they have been widely used historically and remain common today. This essay asks why rulers use them even when individual offenders are known. It develops a non-informational model of repression, based on the notion of collective responsibility: people may be punished for others' actions in order to give them a stake in one another's obedience. By making one person's safety depend on another person's conduct, the ruler creates incentives for people to police each other.

Collective responsibility is modeled as a directed graph, specifying who is punished for whose disobedience. The ruler faces a trade-off between the *vertical deterrence* created by his own sanctions and the *horizontal deterrence* from peer sanctions. Responsibility for others weakens vertical deterrence because obedience no longer guarantees safety from punishment,

1. Humans display a moral aversion to collective punishments from early childhood (Smith and Warneken, 2016; Thomas et al., 2024).

but it can generate horizontal deterrence by inducing people to police those whose disobedience exposes them to punishment. Making people responsible for others increases obedience precisely when the horizontal-deterrence gains exceed the vertical-deterrence losses.

The military of ancient Rome used various forms of collective discipline, which in extreme cases included decimation, the killing by lot of one in ten soldiers. In the dissertation's opening epigraph, Livy recounts the earliest case of decimation in Roman written tradition. After a defeat against the Volscians, consul Appius Claudius beheaded every deserter; then, the rest of the army was decimated. Ancient authors like Polybius (6.38.1–4), as well as modern ones like Eckstein (2005), have suggested that the threat of these practices was important in enforcing discipline among soldiers.[2] Nonlethal collective discipline survives in the modern military: for example, Gilham (1982, p. 239) reports an example of a United States Marine drill instructor using collective punishment, while reminding everyone that offending recruits may always "accidentally" injure themselves or slip in the shower.

In premodern China, extreme forms of collective punishment were common, including the extermination of the families of those accused of high crimes (Fairbank, 1978). Shang Yang, the Legalist reformer credited with instituting mutual-responsibility groups of five or ten households (*shíwǔ*), was later accused of treason and, according to tradition, executed along with his clan.[3] The persecution of the families of political opponents has persisted into the modern era, including *Sippenhaft* in Nazi Germany (Loeffel, 2012), Stalin's extermination of the *rod* (kin) of the "enemies of the state" (Alexopoulos, 2008), and the reported confinement of entire families to North Korean labor camps (Haggard and Noland, 2012).

A more recent case comes from the essay's opening quote. In the Gulag, prisoners were often divided into work brigades, with punishment and rewards being based on a group point system (Applebaum, 2003). The stated goal was to ensure that inmates would be compelled by their peers to work hard and obey the rules, for example through beatings and

2. The intensity of collective discipline varied: executions were rare in practice though always a staple of military law, while nonlethal punishments were far more common (Taylor, 2022).

3. The next essay discusses *shíwǔ* and related denunciation-based institutions in greater depth.

expulsion into weaker brigades with reduced rations. Pyotr Yakir (1972, p. 100), a historian who spent his youth in the Gulag, recalls that after his failed escape, his colony was docked many points; another inmate warned, "[if] you go into the compound now, the lads will tear you apart." Less extreme forms still exist in modern prisons, despite formal prohibitions (McCall-Smith, 2016; Shaw et al., 2023). Heckathorn (1988, p. 537) cites a juvenile officer in the U.S. who notes that revoking privileges for the whole block after a fight makes repeat offenders "very unpopular with their peers."

Despite their frequency, the precise structure of incentives generated by collective punishments remains understudied. Individual responsibility makes good rational sense, because the deterrent effect of a sanction is largest when the expected gap in punishment between disobeying and obeying is maximized (e.g., Becker, 1968). Perhaps because of this, scholars have typically attributed collective punishments to informational constraints on the part of the authority, who uses group-based sanctions if unable to reliably observe and punish behavior individually (e.g., Kalyvas, 2006).

However, in all examples above the informational account is incomplete, as it does not explain why known non-offenders are deliberately sanctioned after the offender has been identified. Why would Claudius first execute the identified deserters and only then decimate the remaining soldiers? Why persecute the families of political enemies if they are not accused of any crime? Why punish Yakir's brigade members, who had stayed behind, for his escape? The puzzle is especially stark in disciplinary institutions such as armies, prisons, and labor camps (Goffman, 1958; Foucault, 1977): these are settings in which individual behavior is closely scrutinized, yet the punishment of the innocent remains pervasive.

2.1.1 Summary of results

A ruler maximizes obedience over a finite set N of people by publicly designing a *collective responsibility* regime, represented by a relation $R \subseteq N \times N$. The ordered pair iRj means that person i is punished when person j disobeys. People observe R and publicly decide

whom to police. If i polices j and j disobeys, then j is sanctioned and i pays a cost. People then privately learn their types and publicly decide whether to obey.

Fixing a responsibility relation and a policing profile, each person's obedience rule is a cutoff in her type. Vertical deterrence comes from the extent to which her own obedience protects her from ruler punishment. Responsibility for others weakens this channel, because she may be punished even if she obeys. Horizontal deterrence comes from peer sanctions: responsibility can induce people to police those whose obedience reduces their own exposure to punishment. Section 2.2 isolates this trade-off in the two-person case. Theorem 2.4 derives the obedience equilibria and comparative statics in the general model.

The general model then characterizes how horizontal enforcement propagates through a responsibility network. Theorem 2.6 shows that equilibrium policing follows a *connection* relation: a person polices another only when the other's obedience can reduce her own exposure to ruler punishment. When policing is cheap, every such connection induces policing.

When sanctioning peers is too costly to incentivize, individual responsibility is uniquely optimal (Proposition 2.9). When sanctioning others is cheap, two main results follow. First, reaching one person's maximal obedience cutoff requires preserving that person's full vertical deterrence while concentrating all possible peer sanctions on her, so maximal obedience cannot be achieved for more than one person at a time (Proposition 2.12). Second, the optimal symmetric responsibility network is maximally sparse, and the ruler punishes individually if and only if peer sanctions are small (Proposition 2.10).

Section 2.4 considers several additional mechanisms. Mutual responsibility can make policing a positive signal of the policer's own obedience. Without ex ante commitment, peer sanctions must be credible after disobedience is observed. With altruistic concerns, people may obey or police in order to reduce punishment borne by those they care about.

All proofs for this essay and the two that follow are in the Appendix.

2.1.2 Related literature

This essay contributes to research on collective sanctions, repression, and political violence.

The closest antecedents on collective sanctions and in-group enforcement are arguably in sociology (Heckathorn, 1988, 1990; Whitmeyer, 2002) and legal theory (Levinson, 2003), where the present mechanism has been hypothesized without making the strategic trade-off and the ruler's design problem explicit.

Political economists often view group incentives, reputations, and peer enforcement as a second-best response to imperfect observability of individual behavior (e.g., Tirole, 1996; Greif, 2002). In Fearon and Laitin (1996), individual histories are common knowledge within but not across groups. Repression is also usually studied as an informational problem (e.g., Egorov and Sonin, 2024). Rulers use purges to motivate and screen unobservable agents (Montagnes and Wolton, 2019), vary repression by group when information is unreliable (Rozenas, 2020), or choose whether to punish rebels alone or also silent free-riders when rebels trade off group size and secrecy (Zhou, 2021). The essay argues that group-based incentives, and collective repression in particular, can be optimal even when such informational constraints are absent.

At a general level, we can say that collective punishment exposes people to sanctions due to others' actions, while indiscriminate violence is imposed without regard to behavior. However, the literature often blurs these concepts, and treats them as synonymous with targeting civilians or non-combatants.[4] Much of this literature argues that indiscriminate violence is counterproductive, producing backlash and mobilizing opposition (Lichbach, 1987; Rosendorff and Sandler, 2010; Hultquist, 2017; Kocher et al., 2011). Other work finds more conditional effects: Lyall (2009), for example, finds that indiscriminate shelling of Chechen villages reduced subsequent insurgent attacks.

4. Kalyvas (2006, p. 142) writes: "The distinction is based on the level at which 'guilt' (and hence targeting) is determined. Violence is selective when there is an intention to ascertain individual guilt [...] selective violence entails personalized targeting, whereas indiscriminate violence implies collective targeting [...] individual guilt is replaced by the concept of guilt by association."

The distinction between indiscriminate violence and collective punishment matters in both directions. Indiscriminate violence may be aimed at non-punitive ends, such as draining enemy support, destroying infrastructure, trimming the population of ideological enemies, appropriating resources, or manipulating the labor market (Valentino et al., 2004; Lyall, 2009; Gregory et al., 2011; Esteban et al., 2015; Sun, 2024).[5] Conversely, practices that have been described as selective violence are obvious examples of collective punishment. Punitive demolitions of insurgents' houses have been used by the Israel Defense Forces in the West Bank and by pro-Russian security forces in Chechnya (Darcy, 2007; Klocker, 2020). Benmelech et al. (2015) find that punitive demolitions of Palestinian suicide bombers' houses reduced further attacks in the short term during the Second Intifada.[6] Although they frame their results as evidence of the effectiveness of selective counter-insurgency (in contrast with the indiscriminate bulldozing of whole city blocks), punitive house demolitions are collective punishment because an entire household is sanctioned for one person's actions.[7] Souleimanov et al. (2022) make a similar point with regard to punitive kin killings.

This essay also contributes to work on agency and credibility in repression. Tyson (2018) argues that repression is often modeled as an interaction between ruler and targets, ignoring agency problems between ruler and apparatus. Acemoglu and Wolitzky (2020) study the trade-off between community enforcement and specialized enforcement in repeated games, emphasizing that specialized enforcers must themselves be given incentives to punish. In this essay, the ruler does not choose between specialized and community enforcement, but rather designs responsibility relations that make targets of repression responsible for enforcing oth-

5. The law of war makes a similar distinction. Although collective punishment may involve group-based targeting, it requires the specific *mens rea* of punishing a particular outcome. According to the Sierra Leone Tribunal, "While targeting takes place on account of who the victims are [. . .] the crime of collective punishments occurs in response to the acts or omissions of protected persons" (Klocker, 2020, p. 30).

6. Fallah (2024) questions the robustness of their results. Hatz (2019) shows that Palestinians view punitive demolitions as collective punishments and that they increase opposition to Israel.

7. The logic of family responsibility appears explicitly in the Supreme Court of Israel's house-demolition jurisprudence. Justice Sohlberg, invoking the principle that "the sinner is not alone," attributed responsibility for a relative's act to the whole family, reasoning that they "should have known" of it and "should have done something to prevent the act" (Klocker, 2020, pp. 55–56).

ers' obedience. Bueno de Mesquita et al. (2024) show that when repression is costly for both government and citizens, too large a threat may lack credibility and increase opposition. In this model, the ruler's trade-offs do not depend on any punishment cost.[8]

Group-liability lending, historically associated with microcredit institutions like Grameen Bank, relied on collective sanctions: groups of borrowers were jointly responsible for repayment, and each person's future loans could depend on all members' good standing (Morduch, 2000). Grameen Bank later moved away from formal joint liability, though the mechanism remains central to classic theories of group lending and peer monitoring (Stiglitz, 1990; Varian, 1990). The evidence of the effectiveness of group-liability lending is mixed (Giné and Karlan, 2014), and so is the lab-experimental evidence of collective sanctions more broadly (e.g., Dickson, 2007; Pereira and van Prooijen, 2018; Chapkovski, 2021; Duell et al., 2024).

2.2 Motivating example with two people

This section isolates the trade-off between vertical and horizontal deterrence in the smallest environment in which both margins are present.

2.2.1 Model setup

There are three actors: a ruler, Ann, and Bob. Ann and Bob each choose whether to obey, $a_i \in \{0, 1\}$, where $a_i = 0$ denotes obedience and $a_i = 1$ disobedience.

Ann and Bob have private gains from disobedience θ_A and θ_B, independently drawn from a common distribution F on $\mathbb{R}$. F is strictly increasing and continuous, and $\mathbb{E}[|\theta_i|] < \infty$. Each person observes her own type before choosing whether to obey.

The ruler first chooses whether to make Ann responsible for Bob, $r \in \{0, 1\}$. If $r = 1$, Ann is punished if at least one of Ann and Bob disobeys; if $r = 0$, she is punished only when

8. A separate class of theories studies *signaling*. Terrorists may provoke indiscriminate violence to reveal government disregard for civilians (Bueno de Mesquita and Dickson, 2007); rulers may use repression to signal loyalty of the apparatus (Casper and Tyson, 2014); and indiscriminate repression may signal inability to monitor (Rozenas, 2020).

she disobeys. Bob is punished whenever he disobeys. Punishment by the ruler imposes a unit loss on its recipient.

After observing r, but before types are realized, Ann publicly chooses whether to police Bob, $P \in \{0,1\}$. If she polices Bob and he disobeys, Bob incurs a peer sanction γ and Ann incurs a policing cost ξ. Types are then realized, and Ann and Bob simultaneously choose whether to obey after observing their own types. This timing represents an ex ante commitment to sanctioning behavior; Section 2.4.2 considers variants without commitment. The one-directional structure is expositional; the general model allows arbitrary directed responsibility and policing relations.

The ruler values obedience according to $V(a_A, a_B) = (1-\beta)(1-a_A) + \beta(1-a_B)$, where $\beta \in (0,1)$ is the weight on Bob's obedience.

Ann's and Bob's payoffs are

$$u_A(a_A, a_B; r, P) = \theta_A a_A - \Phi_A(a_A, a_B; r) - \xi P a_B,$$
$$u_B(a_B; P) = a_B(\theta_B - 1 - P\gamma),$$

where $\xi, \gamma > 0$ and $\Phi_A(a_A, a_B; r) := a_A + (1-a_A) r a_B$ indicates whether Ann is punished. Throughout the two-person model, each person obeys when indifferent and Ann chooses $P = 1$ when indifferent.

2.2.2 Obedience decisions

Fix $(r, P) \in \{0,1\}^2$. The obedience stage has a unique cutoff equilibrium. Bob obeys if and only if $\theta_B \leq 1 + P\gamma$, so his cutoff is $x_B(P) = 1 + P\gamma$.

Ann is punished after obeying only when $r = 1$ and Bob disobeys, whereas disobedience leads to punishment for sure. Since Bob obeys with probability $F(x_B(P))$, Ann obeys if and only if

$$\theta_A \leq x_A(r, P) := 1 - r\Big(1 - F(x_B(P))\Big) = 1 - r\Big(1 - F(1 + P\gamma)\Big). \tag{2.1}$$

Policing raises Bob's cutoff from 1 to $1+\gamma$. Under individual responsibility, Ann's cutoff remains 1. Under collective responsibility, her cutoff is $F(1)$ without policing and $F(1+\gamma)$ with policing. Holding policing fixed, responsibility for Bob weakens Ann's incentive to obey because her own obedience no longer guarantees protection from punishment. Policing partly restores that incentive by increasing Bob's obedience.

2.2.3 Ann's policing decision

Define $J(x) := \int_1^x F(t)\,dt$. By Lemma A.1, $J(x)$ is the change in Ann's expected obedience-stage payoff when her cutoff moves from 1 to x. Up to an additive constant independent of (r, P), Ann's ex ante continuation payoff is

$$W_A(r,P) = J\Big(x_A(r,P)\Big) - \xi P\Big(1 - F(1+\gamma)\Big). \tag{2.2}$$

Under individual responsibility, policing is never worthwhile. If $r = 0$, then $x_A(0,1) = x_A(0,0) = 1$ and $W_A(0,1) - W_A(0,0) = -\xi\Big(1 - F(1+\gamma)\Big) < 0$.

Under collective responsibility, policing raises Bob's cutoff from 1 to $1+\gamma$ and Ann's cutoff from $F(1)$ to $F(1+\gamma)$. Ann polices if and only if $J\Big(F(1+\gamma)\Big) - J\Big(F(1)\Big) \geq \xi\Big(1 - F(1+\gamma)\Big)$, or equivalently,

$$\xi \leq \frac{J\Big(F(1+\gamma)\Big) - J\Big(F(1)\Big)}{1 - F(1+\gamma)}. \tag{2.3}$$

The numerator is Ann's gross gain from policing, and the denominator is the probability that she incurs the policing cost. The threshold is strictly positive for every $\gamma > 0$ and converges to zero as $\gamma \to 0$.

2.2.4 The ruler's problem

The ruler benefits from collective responsibility only if it induces policing.

Under individual responsibility, Ann does not police and both people obey with probability $F(1)$, so the ruler's payoff is $V_0 = F(1)$. If he chooses collective responsibility

but (2.3) fails, Bob still obeys with probability $F(1)$, whereas Ann obeys with probability $F(F(1)) < F(1)$. Collective responsibility without policing is therefore strictly worse than individual responsibility.

Suppose instead that (2.3) holds. Ann and Bob then obey with probabilities $q_A^1 := F\Big(F(1+\gamma)\Big)$ and $q_B^1 := F(1+\gamma)$, where the superscript denotes collective responsibility and $q_A^1 < F(1) < q_B^1$. The ruler's payoff is $V_1 = (1-\beta)q_A^1 + \beta q_B^1$. Define

$$\beta^*(\gamma) := \frac{F(1) - F\Big(F(1+\gamma)\Big)}{F(1+\gamma) - F\Big(F(1+\gamma)\Big)}. \tag{2.4}$$

This threshold lies in $(0,1)$ because $q_A^1 < F(1) < q_B^1$.

Proposition 2.1 (Optimal binary responsibility)**.** *If* (2.3) *fails, individual responsibility is uniquely optimal. If* (2.3) *holds, the ruler weakly prefers collective responsibility if and only if* $\beta \geq \beta^*(\gamma)$ *and strictly prefers it if* $\beta > \beta^*(\gamma)$*.*

At $\beta^*(\gamma)$, the weighted gain in Bob's obedience exactly equals the weighted loss in Ann's obedience.

2.2.5 Continuous responsibility

Suppose now that the ruler chooses a responsibility level $r \in [0,1]$, retaining the same objective as above. Ann is punished for sure if she disobeys and with probability r if she obeys while Bob disobeys. Thus, $r = 0$ represents individual responsibility and $r = 1$ full responsibility.[9]

Bob's cutoff remains $x_B(P) = 1 + P\gamma$, while Ann's becomes

$$x_A(r,P) = 1 - r\Big(1 - F(1+P\gamma)\Big). \tag{2.5}$$

9. $r = \frac{1}{10}$ captures the individual punishment risk created by decimation: following a collective failure, one tenth of the unit was selected by lot. The analogy is reduced-form, since the model represents individual liability as a probability rather than the joint selection of a fixed fraction of the group.

Up to an additive constant independent of (r, P), her continuation payoff remains

$$W_A(r, P) = J\Big(x_A(r, P)\Big) - \xi P\Big(1 - F(1 + \gamma)\Big). \tag{2.6}$$

Ann's gross gain from policing is $\mathscr{G}(r; \gamma) := J\Big(x_A(r, 1)\Big) - J\Big(x_A(r, 0)\Big)$. She polices if and only if $\mathscr{G}(r; \gamma) \geq \xi\Big(1 - F(1 + \gamma)\Big)$.

The function $r \mapsto \mathscr{G}(r; \gamma)$ need not be monotone. Increasing r widens the gap between Ann's cutoffs with and without policing, but also lowers both cutoffs. Because $J'(x) = F(x)$, a given increase in the cutoff is less valuable at lower cutoff levels. Define

$$\mathcal{X}(\xi, \gamma) := \Big\{r \in [0, 1] : \mathscr{G}(r; \gamma) \geq \xi\Big(1 - F(1 + \gamma)\Big)\Big\}. \tag{2.7}$$

Because $\mathscr{G}(0; \gamma) = 0$, individual responsibility does not induce policing. Moreover, F is continuous, so $\mathcal{X}(\xi, \gamma)$ is closed and has a strictly positive minimum whenever it is nonempty.

Proposition 2.2 (Minimal responsibility)**.** *If $\mathcal{X}(\xi, \gamma)$ is empty, individual responsibility is uniquely optimal. Otherwise, let $\underline{r} := \min \mathcal{X}(\xi, \gamma) > 0$. Every optimal responsibility level belongs to $\{0, \underline{r}\}$.*

Any positive responsibility level that fails to induce policing lowers Ann's obedience without affecting Bob's and is dominated by $r = 0$. Among levels that induce policing, the ruler chooses the smallest because further responsibility reduces Ann's obedience without further increasing Bob's.

2.3 Model of collective responsibility

The two-person example made collective responsibility a single choice: whether Ann is responsible for Bob. The general model represents responsibility and policing as relations on

the same population.[10] This allows the analysis to compare sparse and dense, symmetric and asymmetric structures while preserving the central trade-off.

2.3.1 Model setup

Let $N = \{1, \ldots, n\}$ be a finite set of people, where $n \geq 2$. Each person $i \in N$ chooses whether to obey ($a_i = 0$) or disobey ($a_i = 1$). An action profile is $a = (a_i)_{i \in N} \in \{0, 1\}^n$. Each person i has a privately observed gain from disobedience, $\theta_i \in \mathbb{R}$, drawn independently from a common distribution F. Assume that F is continuous and strictly increasing on the real line and that $\mathbb{E}[|\theta_i|] < \infty$.[11]

The ruler chooses a *responsibility relation* $R \subseteq N \times N$, where iRj means that person i is punished whenever person j disobeys. Person i is punished if at least one person for whom she is responsible disobeys. Thus

$$\Phi_i(R, a) = \max_{j : iRj} a_j, \tag{2.8}$$

where the maximum over the empty set equals zero. Being punished by the ruler costs $\kappa > 0$.

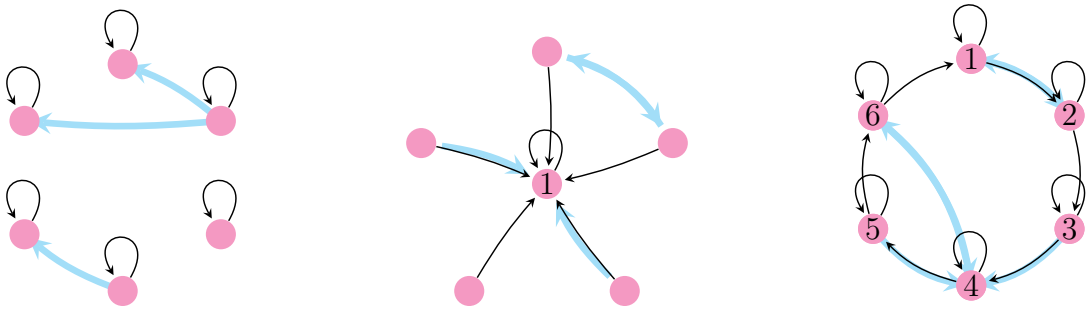
Each person i also chooses a set $P_i \subseteq N \setminus \{i\}$ of people whom she polices. These choices define an irreflexive relation $P \subseteq N \times N$, where iPj means that i polices j. If iPj and j disobeys, then i incurs a policing cost $\xi > 0$ and j incurs a peer sanction $\gamma > 0$. Let

$$d_i(P) := |\{j \in N : jPi\}|$$

denote the number of people who police i. Responsibility and policing are therefore directed relations on the same population: the ruler chooses R, while P emerges from people's policing choices under it.

10. The binary model in Section 2.2 is the special case $n = 2$ and $\kappa = 1$, with responsibility restricted to I and $I \cup \{(A, B)\}$, policing restricted to Ann policing Bob, and a linear ruler objective with weights $(1-\beta, \beta)$.

11. The moment condition ensures that the expected continuation payoffs used below are finite.



(a) Individual responsibility $I := \{(i,i) : i \in N\}$

(b) Star responsibility $S_1 := \{(i,1) : i \in N\}$

(c) Cycle responsibility C

Figure 2.1. Responsibility and policing as graphs on the same vertex set of people. Thin black arrows represent responsibility R; thick blue arrows represent policing P.

An arrow $i \to j$ in the responsibility graph indicates that i is punished when j disobeys. Arrows therefore point from the person exposed to punishment toward the person whose disobedience triggers it, and self-responsibility is represented by a loop. Three responsibility rules will recur throughout the analysis. Under *individual responsibility*, $I := \{(i,i) : i \in N\}$, each person is responsible only for herself. Under the *star* centered on i, $S_i := \{(j,i) : j \in N\}$, everyone is responsible for person i. Finally, under *cycle responsibility*,

$$iCj \quad \text{if and only if} \quad i = j \quad \text{or} \quad j = i + 1 \pmod{n},$$

each person is responsible for herself and one neighbor, forming a directed cycle. A cycle responsibility rule is any relabeling of this cycle. Figure 2.1 illustrates these responsibility rules together with illustrative policing relations.

The three rules preview the main design possibilities studied below: preserving individual deterrence, concentrating peer enforcement on one person, and distributing it symmetrically through a sparse network.

The timing is illustrated in Figure 2.2. First, the ruler publicly chooses R. Second, people simultaneously and publicly choose whom to police. Third, nature independently draws $(\theta_i)_{i \in N}$ from F. Finally, each person privately observes her type and chooses whether to obey.

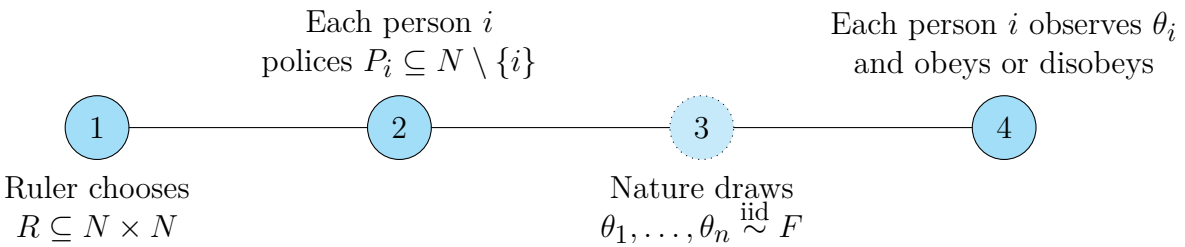


Figure 2.2. Timing of the general model of collective responsibility. All actions are publicly observed, but people learn their types privately.

In the baseline, people choose whom to police before learning their gains from disobedience. Their policing decisions therefore reflect the incentives created by the responsibility relation rather than their privately known temptations to disobey. Section 2.4.1 allows types to be realized before policing decisions, while Section 2.4.2 considers variants in which peer sanctions must be credible after disobedience is observed.[12]

The ruler maximizes obedience: his payoff is $v(a)$, where $v : \{0,1\}^n \to \mathbb{R}$ is strictly decreasing in each coordinate. Changing any person's action from disobedience to obedience, holding all other actions fixed, strictly raises his payoff. Each person i's utility is

$$U_i(R, P, a) = \theta_i a_i - \kappa \Phi_i(R, a) - \xi \sum_{j : iPj} a_j - \gamma d_i(P) a_i. \tag{2.9}$$

The four terms are i's gain from disobedience, ruler punishment, the cost of policing others who disobey, and the peer sanctions imposed on i when she disobeys. The analysis proceeds by backward induction, beginning with obedience decisions, then policing decisions, and finally the ruler's design.

12. Policing costs are paid only after disobedience: if i polices j and j disobeys, then i pays ξ and j bears the sanction γ. Alternatively, policing could require an up-front monitoring cost, with the peer sanction triggered costlessly by disobedience. This would increase the cost threshold for activating a policing link while leaving the obedience cutoffs and connection relation unchanged.

2.3.2 Obedience decisions

Fix a responsibility relation R and a policing relation P. In the obedience subgame, each person i privately observes θ_i and chooses $a_i \in \{0, 1\}$. A person is assumed to obey when indifferent. The payoff difference between disobeying and obeying is increasing in θ_i. Hence every equilibrium of the obedience subgame is in cutoff strategies: for each $i \in N$, there exists $x_i \geq 0$ such that person i obeys if and only if $\theta_i \leq x_i$.

The cutoff x_i is the payoff gain from obeying. It has two parts. The first is ruler punishment avoided by obeying. This term is present only if i is self-responsible: obeying prevents i from triggering her own punishment only when every other person for whom she is responsible obeys. The second is the peer sanction avoided by obeying. If jPi, then i incurs a sanction γ when she disobeys. Thus equilibrium cutoffs satisfy

$$x_i = \kappa \mathbf{1}\{iRi\} \prod_{\substack{j \neq i \\ iRj}} F(x_j) + \gamma d_i(P) =: g_i(x). \tag{2.10}$$

The first term is vertical deterrence; the second is horizontal deterrence.

Self-responsibility performs two roles: it makes a person's obedience pivotal for her own punishment, creating vertical deterrence, and also makes her obedience depend on the conduct of those for whom she is responsible. Self-responsible rows therefore transmit changes in obedience through the responsibility relation; rows without self-responsibility do not.

Denote the largest attainable equilibrium obedience cutoff by $\overline{\theta} := \kappa + (n-1)\gamma$.

Remark 2.3. Fix R and P. Every equilibrium of the obedience subgame is in cutoff strategies. Moreover, $x \in [0, \overline{\theta}]^n$ is an equilibrium cutoff profile if and only if it satisfies (2.10).

To identify the part of the responsibility relation through which obedience incentives propagate, define

$$\overline{R} := \{(i, j) \in R : iRi\}. \tag{2.11}$$

For any $i \neq j$, a higher cutoff for j raises i's cutoff if and only if $i\overline{R}j$. Thus $\overline{R}$ consists precisely of the responsibility links that transmit changes in obedience.

The comparative statics follow from the cutoff map g in (2.10). Increasing κ raises g_i for every self-responsible i. Increasing γ, or adding a policing link into i, raises g_i directly. Adding a self-responsibility link (i, i) also raises g_i, since it creates the ruler-punishment term in (2.10). Holding self-responsibility fixed, adding an off-diagonal responsibility link (i, j) weakly lowers g_i: if i is not self-responsible, the link is irrelevant; if i is self-responsible, it adds another factor $F(x_j) \leq 1$ to the product that determines the punishment avoided by obeying.

Theorem 2.4 (Obedience comparative statics)**.** *The obedience game has smallest and largest equilibrium cutoff profiles, denoted $x^{\min}$ and $x^{\max}$. Both are weakly increasing in κ and γ. With relations ordered by set inclusion, both are weakly increasing in P and, holding $R \setminus I$ fixed, in $R \cap I$; holding $R \cap I$ fixed, both are weakly decreasing in $R \setminus I$.*

The proof is a standard monotone fixed-point argument. For fixed primitives, g is increasing in x, so Tarski's theorem gives smallest and largest fixed points; monotone comparative statics for extremal fixed points then gives the result.

Because obedience decisions are strategic complements along $\overline{R}$, the obedience subgame may have multiple equilibria. In what follows, either the smallest or the largest equilibrium is fixed, and the same selection is used throughout.

Off-diagonal responsibility therefore has a direct cost for the ruler: holding policing fixed, it weakly lowers obedience. Collective responsibility can be optimal only if the policing it induces more than offsets this loss. The analysis now turns to that policing channel.

2.3.3 Policing decisions

This subsection analyzes the policing decisions induced by R. Let $x^*(R, P)$ denote the selected extremal obedience equilibrium fixed above.

Self-responsible rows transmit changes in obedience through the responsibility relation. A person may therefore have an incentive to police someone for whom she is not directly responsible: that person's obedience may affect an intermediate self-responsible person and thereby change her own exposure to ruler punishment. Responsibility can thus create both direct and indirect policing incentives.

Given responsibility R and cutoff vector x, write $a \sim x$ when actions are independent and $\Pr(a_i = 0) = F(x_i)$ for every i. Define

$$\overline{\Phi}_i(R, P) := \mathbb{E}_{a \sim x^*(R,P)} \Phi_i(R, a) = 1 - \prod_{j : iRj} F\Big(x_j^*(R, P)\Big)$$

as the selected equilibrium probability that i is punished by the ruler.

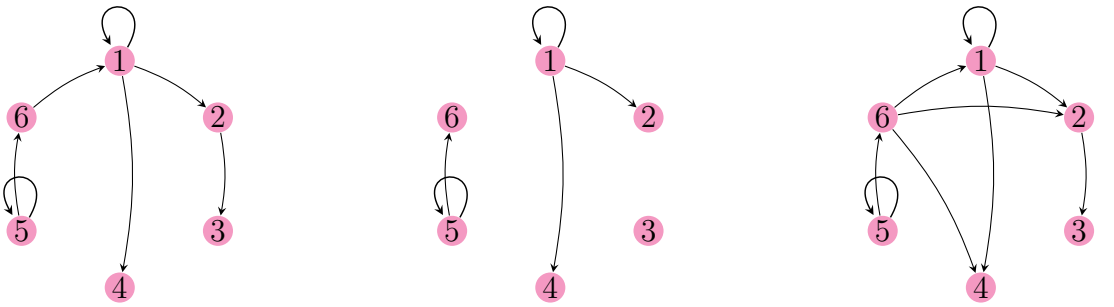
The *induced policing game* is the normal-form game in which people simultaneously choose their outgoing links in P, holding R, κ, γ, ξ fixed. Let $V_i(R, P; \kappa, \gamma)$ denote person i's selected normalized continuation payoff from the obedience subgame, evaluated with outgoing policing costs set equal to zero.[13] Expected payoffs in the induced policing game are

$$\widetilde{U}_i^R(P; \kappa, \gamma, \xi) := V_i(R, P; \kappa, \gamma) - \xi \sum_{j : iPj} \Big(1 - F(x_j^*(R, P))\Big). \tag{2.12}$$

Figure 2.3 illustrates how responsibility paths create indirect policing incentives. Person 6 is directly responsible for person 1. Because person 1 is self-responsible and responsible for persons 2 and 4, greater obedience by either 2 or 4 raises person 1's obedience and thereby lowers person 6's punishment risk. Person 6 is therefore connected to persons 1, 2, and 4.

Definition 2.5. Fix responsibility R. Person i is *connected* to j, denoted $i \xrightarrow{R} j$, if there exists ℓ such that $iR\ell$ and either $\ell = j$ or j is reachable from ℓ via $\overline{R}$. Equivalently, $\xrightarrow{R} := R \circ (I \cup \overline{R}^+)$, where composition is read from left to right.

13. The normalization subtracts the constant $\mathbb{E}[\theta_i]$ and therefore does not affect the policing game.



(a) An arbitrary relation R. (b) The restriction $\overline{R}$. (c) The connection $\overrightarrow{R}$.

Figure 2.3. Construction of the connection relation, starting from an arbitrary relation R (left), through its restriction $\overline{R}$ (center) and finally its connection $\overrightarrow{R}$ (right).

Say that a policing relation P *can arise in equilibrium* if some mixed Nash equilibrium of the induced policing game assigns positive probability to P. The next result makes the relationship between connection and policing exact.

Theorem 2.6 (Connection and policing)**.** *For every $\kappa, \gamma > 0$, there exist $0 < \underline{\xi}(\kappa, \gamma) \leq \overline{\xi}(\kappa, \gamma) < \infty$ such that, for every $\xi > 0$, every responsibility relation R, every policing relation P that can arise in equilibrium, and every pair $i \neq j$:*

(a) if iPj, then $i \overset{R}{\to} j$ and $\xi \leq \overline{\xi}(\kappa, \gamma)$;

(b) if $\xi \leq \underline{\xi}(\kappa, \gamma)$, then iPj if and only if $i \overset{R}{\to} j$.

The proof builds on the obedience comparative statics. If i is not connected to j, policing j cannot lower i's exposure to ruler punishment. Any collection of unconnected outgoing policing links is therefore strictly dominated by deleting all of them. If i is connected to j, policing j raises j's cutoff directly by γ and, by Theorem 2.4, weakly raises the selected extremal cutoff profile. Strict monotonicity of F, together with $F(0) > 0$, propagates the initial strict increase backward along every relevant $\overline{R}$-path, strictly lowering i's punishment probability. Because there are finitely many responsibility and policing relations, this benefit exceeds the policing cost uniformly when ξ is sufficiently small.

The theorem identifies three regions. Policing always follows connection: people police only those whose obedience can affect their own punishment risk. When $\xi \leq \underline{\xi}(\kappa, \gamma)$,

the unique equilibrium policing relation contains exactly the connections between distinct people. When $\xi > \overline{\xi}(\kappa, \gamma)$, the empty relation is the unique equilibrium policing outcome. Between the two bounds, connection remains necessary but need not be sufficient, and multiple policing equilibria may exist.

The result predicts the direction of peer enforcement, as well as its aggregate extent. Peer sanctions should track the exposure created by the responsibility rule: people should police those whose conduct can ultimately affect their own punishment. Indirect enforcement can travel only through self-responsible intermediaries. A responsibility chain ceases to transmit policing incentives when it reaches a person who is not responsible for herself. Very low policing costs activate every such connection between distinct people, whereas sufficiently high costs eliminate peer enforcement.

Connection identifies which policing links can arise. To discipline equilibrium behavior when policing costs are intermediate, the next result imposes a sufficient condition under which policing choices are strategic complements.

Definition 2.7 (Cutoff convexity)**.** The distribution F is *cutoff-convex* if it is twice differentiable and satisfies $F''(x) \geq 0$ for every $x \in [0, \overline{\theta}]$.

Equivalently, cutoff convexity requires the density to be weakly increasing over the feasible range of obedience cutoffs. It holds, for example, for normal and logistic distributions whose location parameter is weakly above $\overline{\theta}$. Cutoff convexity is sufficient for supermodularity; supermodularity may also hold when cutoff convexity fails. Comparisons across values of κ or γ require cutoff convexity over the union of the corresponding feasible cutoff ranges. Cutoff convexity is used only for supermodularity and the comparative statics that follow; the obedience, connection, and design results do not require it.

The responsibility structure pushes toward complementarity for two reasons. First, policing a target raises that target's obedience, lowering the expected cost of additional policing directed at her. Second, punishment exposure is multiplicative, so the marginal value of one person's obedience is greater when the others whose conduct determines punishment are

already more likely to obey. Cutoff convexity ensures that curvature in F does not overturn these forces.

Proposition 2.8 (Supermodular policing)**.** *If F is cutoff-convex, then the induced policing game is supermodular. Hence it has least and greatest pure equilibria, $P^{\min}(R, \kappa, \gamma, \xi)$ and $P^{\max}(R, \kappa, \gamma, \xi)$, and both are weakly increasing in κ and γ.*

The proposition implies that ruler punishment and peer enforcement reinforce one another. A larger peer sanction makes policing more effective, while a larger ruler sanction gives people more to lose from the disobedience of those to whom they are connected. Under cutoff convexity, stronger ruler punishment and stronger peer sanctions therefore expand both extremal policing equilibria. Empirically, the severity of state punishment and the extent of peer enforcement should covary positively, holding the responsibility rule and policing costs fixed.

The same monotonicity need not hold in ξ. A higher ξ raises the cost of policing, but an additional policing link can also raise obedience among people already policed, lowering the expected cost of existing policing links.

2.3.4 Ruler's problem

The ruler chooses a responsibility relation anticipating both obedience and policing. The analysis first identifies when responsibility for others is counterproductive, then characterizes optimal responsibility when policing is cheap and studies how sanction strength changes the ruler's choice. Finally, it shows how responsibility can concentrate obedience incentives on one person.

Fix a pure-equilibrium selection of the policing game whenever policing is not uniquely determined.[14] Given a responsibility relation R, let $x(R)$ denote the induced obedience vector under the maintained equilibrium selections, and let $q_i(R) := F(x_i(R))$. Given the

14. A pure policing equilibrium exists in the cheap- and expensive-policing regions characterized by Theorem 2.6, and under cutoff convexity by Proposition 2.8.

selected cutoff profile, obedience events are independent because types are independent. The ruler's expected payoff is therefore

$$V(R) := \sum_{a \in \{0,1\}^n} v(a) \prod_{i:a_i=0} q_i(R) \prod_{i:a_i=1} \Big(1 - q_i(R)\Big). \tag{2.13}$$

The first result identifies a general reason for the ruler to use individual responsibility. If responsibility for others generates no peer enforcement, its only effect is to weaken vertical deterrence.

Proposition 2.9 (Collective responsibility requires enforcement)**.** *If a responsibility relation $R \neq I$ induces the empty policing relation under the maintained equilibrium selection, then $V(I) > V(R)$. Consequently, for every $\kappa, \gamma > 0$, individual responsibility is uniquely optimal whenever $\xi > \overline{\xi}(\kappa, \gamma)$. Moreover, for every $\xi, \gamma > 0$, there exists $\underline{\kappa}(\xi, \gamma) > 0$ such that individual responsibility is uniquely optimal whenever $0 < \kappa < \underline{\kappa}(\xi, \gamma)$.*

Under individual responsibility, every person has cutoff κ. Without policing, every cutoff is at most κ. A person who is not self-responsible has cutoff zero, while an off-diagonal link in a self-responsible row multiplies the punishment avoided through obedience by an additional probability strictly below one. Every nonindividual relation therefore weakly lowers all cutoffs and strictly lowers at least one.

The proposition gives collective responsibility a clear behavioral implication. Under the baseline assumption that people care only about their own punishment, responsibility for others can improve obedience only through the peer enforcement it induces. Section 2.4.3 shows that collective punishment may still deter without policing when offenders care about those exposed to punishment. Where neither channel is plausible, its use must rest on considerations omitted from the baseline: retribution, if the ruler values punishment as a response to perceived collective guilt; generalized intimidation, if punishing non-offenders is meant to frighten a wider audience; or administrative convenience, if collective sanctions economize on investigation or individualized enforcement.

The next result characterizes optimal responsibility when policing is cheap. Call a responsibility relation R *connection-minimal* if every off-diagonal responsibility link is necessary for its connection relation: for every $(i,j) \in R \setminus I$, $\overset{R\setminus\{(i,j)\}}{\to} \subsetneq \overset{R}{\to}$.

To state the implication for symmetric rules, a bijection $\varphi : N \to N$ is a *symmetry* of R, denoted $\varphi \in \mathrm{Sym}(R)$, if, for every $i, j \in N$, iRj if and only if $\varphi(i)R\varphi(j)$. The relation R is *symmetric* if, for every $i, j \in N$, some $\varphi \in \mathrm{Sym}(R)$ satisfies $\varphi(i) = j$. Symmetry means that the relation does not "name" anyone. Individual responsibility and cycle responsibility are symmetric; a star is not.

Let $\mathcal{C}$ denote the set of cycle-responsibility relations. For fixed $\kappa, \xi > 0$, let $\Gamma_{\kappa,\xi} := \{\gamma > 0 : \xi \leq \underline{\xi}(\kappa, \gamma)\}$ denote the values of the peer sanction for which policing is cheap.

Proposition 2.10 (Optimal responsibility under cheap policing)**.** *Suppose that* $\xi \leq \underline{\xi}(\kappa, \gamma)$.

(a) Every optimal responsibility relation is reflexive and connection-minimal.

(b) Every optimal symmetric responsibility relation belongs to $\{I\} \cup \mathcal{C}$.

Moreover, fix $\kappa, \xi > 0$. *There exists* $\gamma^* > 0$ *such that, among symmetric responsibility relations and for every* $\gamma \in \Gamma_{\kappa,\xi}$, *individual responsibility is uniquely optimal if* $\gamma < \gamma^*$, *while only cycle-responsibility relations are optimal if* $\gamma > \gamma^*$.

Part (a) follows from the two deterrence channels. Adding a missing self-responsibility link creates vertical deterrence and weakly expands connection. Under cheap policing, it therefore weakly expands policing, weakly raising every cutoff and strictly raising at least one. Conversely, once reflexivity is established, removing an off-diagonal link that does not change connection leaves policing unchanged and strictly raises the cutoff of the person whose row contained it. Such a link cannot appear in an optimal relation.

Optimal collective responsibility consequently has a backbone of individual accountability and a minimal set of cross-person links. Every person remains punishable for her own disobedience, and every link holding her responsible for someone else must create an addi-

tional policing incentive. Once a thinner responsibility relation induces the same enforcement network, further links only increase exposure to punishment.

Part (b) applies this logic to symmetric rules. Any nonindividual symmetric rule assigns at least one off-diagonal responsibility link to every person. A cycle assigns exactly one such link per person. Because every person is self-responsible under a cycle, $\overline{R}$ contains the entire directed cycle, and $\overline{R}^{+}$ connects every person to every other. A cycle therefore induces full horizontal deterrence with the smallest possible dilution of vertical deterrence. Any other nonindividual symmetric rule either uses additional off-diagonal links or generates a smaller connection relation.

The remaining comparison is between individual responsibility and cycle responsibility. Individual responsibility preserves full vertical deterrence and generates no peer sanctions. Cycle responsibility weakens vertical deterrence and induces every person to police every other. The first force dominates when peer sanctions are weak, while the second dominates when they are strong.[15]

These results predict that obedience-oriented collective responsibility should preserve punishment for one's own misconduct and use cross-person liability sparingly. Dense or clique-like responsibility creates no additional enforcement when the same connection relation can be generated by a thinner graph. Historically common clique-like rules may instead reflect limits on who can observe and sanction whom, robustness to inactive members or failed links, the ease of administering fixed groups, or a separate benefit from exposing a broad group to punishment.

Cutoff convexity yields broader comparative statics in sanction strength.

Remark 2.11 (Ruler comparative statics). Suppose that F is cutoff-convex over the cutoff ranges generated by the parameter values being compared, and select either the least or

15. At $\gamma = \gamma^{*}$, the comparison depends on the selected extremal obedience equilibrium. Even when the same extremal selection is used throughout, that equilibrium may change discontinuously with γ as fixed points appear or disappear, so individual and cycle responsibility need not yield the same payoff at the threshold.

the greatest pure policing equilibrium throughout. For every fixed responsibility relation R, $V(R)$ is weakly increasing in κ and in γ. The ruler's maximized payoff is therefore weakly increasing in both parameters. Moreover, $V(I)$ is independent of γ, so, for fixed κ and ξ, the set of values of γ for which individual responsibility is optimal is a lower interval.

Stronger ruler and peer sanctions expand both extremal policing equilibria and increase obedience for any fixed policing relation. The design implication is strongest for γ. Individual responsibility induces no peer policing, so its value does not depend on the peer sanction. Once some collective-responsibility relation strictly outperforms individual responsibility, further increases in γ cannot make individual responsibility optimal again.

The corresponding conclusion does not follow for κ. A larger ruler sanction raises obedience under every fixed responsibility relation, including individual responsibility, and therefore does not generally order the ruler's preferred relations. Equilibrium policing also need not be globally monotone in ξ: a higher policing cost directly discourages policing, while additional policing can increase obedience and reduce the expected cost of existing links. Propositions 2.9 and 2.10 instead characterize the high- and low-cost regions.

Finally, responsibility can concentrate horizontal enforcement on one focal person. Such concentration requires an asymmetric position in the responsibility graph.

Proposition 2.12 (Maximal obedience is exclusive)**.** *Suppose that $\xi \leq \underline{\xi}(\kappa, \gamma)$. For every $i \in N$, some responsibility relation induces $x_i = \overline{\theta}$; for example, the star S_i. However, no responsibility relation can induce $x_i = x_j = \overline{\theta}$ for distinct $i, j \in N$.*

Thus at most one person can receive maximal obedience incentives. Attaining $x_i = \overline{\theta}$ requires person i to be self-responsible, responsible for no one else, and policed by every other person. The star S_i is one particularly stark relation satisfying these conditions: it gives person i cutoff $\overline{\theta}$ and every other person cutoff zero. Other responsibility relations can maximize i's cutoff while preserving some deterrence for the remaining people, but none can give a second person the same maximal combination of vertical and horizontal deterrence.

The cases of the extermination of Shang Yang's clan and *Sippenhaft* illustrate how responsibility can be organized around one focal person. Such a star-shaped structure is especially useful when the ruler places disproportionately high value on that person's obedience, as with a prominent political or military role. It may also deter the focal person directly if she cares about the punishment imposed on her relatives; the next section formalizes this second channel, along with several other extensions.

2.4 Robustness checks and extensions

The baseline makes several key assumptions: policing is chosen before types are known and before disobedience occurs; people care only about their own punishment; and the ruler perfectly observes individual conduct. This section relaxes each assumption. Policing can itself signal the policer's propensity to obey, and peer enforcement can survive without advance commitment, although it may collapse when many people disobey at once. When people care about punishment imposed on others, responsibility can strengthen obedience directly as well as induce policing. By contrast, when the ruler's individual signals become almost uninformative, peer enforcement disappears and individual responsibility is uniquely optimal.

2.4.1 Policing as signaling

In the baseline model, policing is chosen before types are realized, so it cannot reveal the policer's type. The analysis now returns to the two-person example (fixing the ruler's punishment at κ) and lets Ann observe θ_A before deciding whether to police Bob. Under mutual responsibility, her decision can then reveal whether she herself is likely to obey.

Ann observes θ_A and chooses $P \in \{0, 1\}$, where $P = 1$ means that she polices Bob. For expositional simplicity, only Ann can police Bob. Bob observes P and his own type θ_B, after which both people choose whether to obey. Given the posterior induced by Ann's strategy

after each observed policing decision, the largest-obedience equilibrium of the continuation game is selected, and Ann polices when indifferent.

Three responsibility relations isolate when this information matters. Under individual responsibility, $R = \{AA, BB\}$, Bob's action has no effect on Ann's punishment, so Ann never polices. Under one-way responsibility, $R = \{AA, AB, BB\}$, Ann is exposed to Bob's disobedience and may therefore police him. Bob's payoff remains independent of Ann's action, so his cutoff is $x_B(P) = \kappa + \gamma P$: policing affects Bob only through her sanction γ.

Under mutual responsibility, $R = \{AA, AB, BA, BB\}$, Bob is also exposed to Ann's disobedience. Let $q_A(P)$ denote Bob's posterior probability that Ann obeys after observing P. Bob's cutoff is $x_B(P) = \kappa q_A(P) + \gamma P$: the term γP is the direct peer sanction, while $\kappa q_A(P)$ captures Bob's inference about Ann's obedience.

A policing decision by i can affect j through information only if jRi, since only then does j care whether i obeys; for the signal to arise from equilibrium policing, i must also care about j's obedience. One-way responsibility can generate horizontal enforcement, while mutual responsibility can also make that enforcement informative.

Assume mutual responsibility. For fixed posterior beliefs $q_A(P)$ and the corresponding continuation equilibrium, write $q_B(P) := F(x_B(P))$. In equilibrium, these beliefs and obedience probabilities must be generated by Ann's policing and obedience strategies. Ann obeys after choosing P if and only if $\theta_A \leq \kappa q_B(P)$. Her gain from policing is

$$G(\theta_A) = (\kappa q_B(1) - \theta_A)^+ - (\kappa q_B(0) - \theta_A)^+ - \xi(1 - q_B(1)). \tag{2.14}$$

Whenever policing raises Bob's obedience probability, $G(\theta_A)$ is weakly decreasing in θ_A, so Ann follows a cutoff strategy in policing, with lower types more likely to police. These types are also more likely to obey and consequently place greater value on Bob's obedience: Bob's disobedience may cause Ann to be punished despite her own obedience. Higher types,

by contrast, are more likely to trigger punishment through their own disobedience, making Bob's action less consequential for them.

The next result gives an equilibrium in which policing perfectly reveals Ann's subsequent obedience.

Proposition 2.13 (Policing as a positive signal)**.** *Suppose* $R = \{AA, AB, BA, BB\}$*. Define*

$$\xi^* := \kappa \frac{F(\kappa+\gamma) - F(\kappa)}{1 - F(\kappa+\gamma)}, \qquad x_P^* := \kappa F(\kappa+\gamma) - \xi(1 - F(\kappa+\gamma)). \tag{2.15}$$

If $0 < \xi < \xi^*$*, then there is a pure-strategy PBE in which Ann polices if and only if* $\theta_A \leq x_P^*$*. In this equilibrium,* $q_A(1) = 1$ *and* $q_A(0) = 0$*, so Bob's cutoff is* $x_B(1) = \kappa + \gamma$ *after policing and* $x_B(0) = 0$ *after no policing.*

Both policing decisions occur with positive probability, so the posterior beliefs $q_A(0)$ and $q_A(1)$ are determined by Bayes' rule. The proposition isolates a fully separating equilibrium and does not characterize all PBEs.

The comparison with one-way responsibility identifies the informational effect. Under one-way responsibility, policing raises Bob's cutoff from κ to $\kappa+\gamma$, whereas in the separating equilibrium under mutual responsibility, it raises his cutoff from 0 to $\kappa+\gamma$. The peer sanction contributes γ, while Ann's inferred obedience contributes κ. No policing reveals that Ann will disobey in the relevant continuation, eliminating the direct incentive that her expected obedience would otherwise create.

2.4.2 Policing without commitment

The baseline timing has people choose policing before obedience decisions, giving policing a commitment interpretation: a person agrees in advance to incur the enforcement cost if the target later disobeys. This section considers two alternatives: Section 2.4.2.1 interprets policing as a contingent switch between self-enforcing continuation equilibria, while

Section 2.4.2.2 removes advance commitment and lets people choose sanctions only after obedience has been observed.

2.4.2.1 Withholding future cooperation

Policing can represent the credible threat to withdraw cooperation in a subsequent bilateral interaction. After obedience decisions, suppose that each ordered pair (i, j) plays the following game:

	C_j	D_j
C_i	(ξ, γ)	$(-\eta, 0)$
D_i	$(0, -\eta)$	$(0, 0)$

with $\xi, \gamma, \eta > 0$. (2.16)

The continuation choices are observed within the group but are neither verifiable nor contractible by the ruler. A switch from CC to DD imposes a loss ξ on person i and a loss γ on person j, matching the policing cost and peer sanction in the baseline model. Both CC and DD are pure-strategy equilibria.

Before obedience decisions, each person i publicly announces, for every $j \neq i$, a map $\pi_{ij} : \{0, 1\} \to \{CC, DD\}$, where $\pi_{ij}(a_j)$ selects the continuation equilibrium played by the ordered pair (i, j) after j chooses a_j. These announcements serve as a public equilibrium-selection device, and following the announced map is sequentially rational because each prescribed outcome is a continuation equilibrium. Attention is restricted to maps satisfying $\pi_{ij}(0) = CC$, so obedience preserves cooperation and withdrawal cannot occur for reasons unrelated to disobedience; the remaining choice is whether disobedience triggers DD.

Remark 2.14. For every responsibility relation R, every such profile of continuation maps induces a unique policing relation P: person i polices person j if and only if their continuation outcome is CC after j obeys and DD after j disobeys. The resulting obedience and policing incentives coincide with those in the baseline model.

This interpretation replaces commitment to a costly future action with a public selection between self-enforcing continuation equilibria. The construction is not renegotiation-proof,

since both people prefer CC to DD after disobedience. The announcements coordinate equilibrium selection; they do not make withdrawal jointly optimal.

2.4.2.2 Ex post policing

Advance commitment is now removed altogether. The ruler first chooses a responsibility relation R; nature then draws types, which people privately observe before choosing whether to obey. Only after the obedience profile is realized does each person choose whom to sanction.

The ruler's punishment rule is modified accordingly. If iRj and $i \neq j$, then i is punished when j disobeys unless i sanctions j; if iRi, then i is punished whenever she disobeys. Because sanctioning another person cannot undo punishment caused by one's own disobedience, person i is punished if and only if some disobedient person for whom she is responsible remains unsanctioned:

$$\widetilde{\Phi}_i(R, a, P) := \max_{j : i(R \setminus P) j} a_j, \tag{2.17}$$

where the maximum over the empty set equals zero. Because P is irreflexive, self-responsibility can never be removed through sanctioning. Sanctions impose cost ξ on the sanctioner and cost γ on the target, as in the baseline model. Payoffs are obtained from (2.9) by replacing Φ_i with $\widetilde{\Phi}_i$.

By backward induction, fix R and an obedience profile a. Let $D_i(R, a) := \{\ell \neq i : iR\ell \text{ and } a_\ell = 1\}$ be the set of disobedient people other than i for whom i is responsible, and let $m_i(R, a) := |D_i(R, a)|$.

Person i can use sanctions to avoid ruler punishment only if she is not already punished by her own disobedience. Define

$$\mathscr{S}(R, a) := \{i \in N : (i, i) \notin R \text{ or } a_i = 0\}.$$

Thus $\mathscr{S}(R,a)$ is the set of people who can still avoid ruler punishment by sanctioning the disobedient people for whom they are responsible.

Proposition 2.15 (Ex post sanctioning)**.** *For every R and a, there exists an equilibrium sanctioning relation P^* such that, for every distinct $i \neq j$, iP^*j if and only if $i \in \mathscr{S}(R,a)$, $j \in D_i(R,a)$, and $\xi m_i(R,a) \leq \kappa$.*

Given the sanctioning relation P^*, every equilibrium of the obedience stage is in cutoff strategies. Because $m_i(R,a)$ does not depend on a_i, write it as $m_i(R,a_{-i})$, and define

$$s_i(R,a_{-i}) := \Big|\{h \neq i : hP^*(R,(1,a_{-i}))i\}\Big|.$$

In any cutoff equilibrium,

$$x_i = \mathbf{1}\{iRi\}\mathbb{E}\left[\kappa - \min\{\xi m_i(R,a_{-i}),\kappa\}\right] + \gamma\mathbb{E}\left[s_i(R,a_{-i})\right], \tag{2.18}$$

where expectations are taken over the independent actions induced by the other people's cutoffs. The first term is the net ruler punishment avoided by obedience; the second is the expected peer sanction avoided.

Ex post enforcement is all or nothing: because partial sanctioning incurs a cost without preventing ruler punishment, a person sanctions everyone in $D_i(R,a)$ or no one. Conditional on still being able to avoid punishment, person i sanctions exactly when the total cost $\xi m_i(R,a)$ does not exceed κ, so her ex post enforcement capacity is $\lfloor \kappa/\xi \rfloor$ simultaneously disobedient people. Once $m_i(R,a)$ exceeds this bound, she sanctions no one; peer enforcement may therefore deter isolated disobedience but collapse when many people disobey simultaneously.

The right-hand side of (2.18) defines an increasing self-map of $[0,\overline{\theta}]^n$. Greater obedience by others lowers $m_i(R,a)$ and can make sanctioning worthwhile; if i is self-responsible, her own obedience determines whether sanctions can still save her from ruler punishment.

Tarski's fixed-point theorem therefore gives smallest and largest cutoff equilibria. The same complementarity operates in both directions: higher obedience makes peer enforcement more likely, while anticipated peer enforcement raises the incentive to obey.

The ex post model also preserves the concentration result from the baseline model. Recall that $\overline{\theta} := \kappa + (n-1)\gamma$ is the largest possible obedience cutoff.

Remark 2.16 (Maximal obedience is again exclusive). Suppose that $\xi \leq \kappa$. For every $i \in N$, some responsibility relation induces $x_i = \overline{\theta}$; for example, the star S_i. No responsibility relation can induce $x_i = x_j = \overline{\theta}$ for distinct $i, j \in N$.

The star S_i gives person i the full direct incentive κ and all $n-1$ peer sanctions whenever she disobeys; each other person is responsible only for i, so each has only one person to sanction and does so when $\xi \leq \kappa$.

The same construction cannot maximize two people simultaneously. A maximally deterred person must be self-responsible; when two self-responsible people both disobey, however, each is already punished and cannot escape punishment by sanctioning the other. Neither sanctions the other in that state, so neither can receive every possible peer sanction whenever she disobeys. The source of exclusivity therefore differs from the baseline model: here it follows from ex post credibility rather than dilution of vertical deterrence.

2.4.3 Altruistic concerns

The baseline model assumes that people care only about their own punishment; the model now allows each person to internalize punishment imposed on others. Let $L \subseteq N \times N$ be an exogenously given irreflexive directed relation. If iLj, then person i incurs cost $\lambda > 0$ whenever person j is punished by the ruler. Given responsibility R, policing P, and action profile a, person i's utility is

$$U_i^L(R, P, a) = U_i(R, P, a) - \lambda \sum_{j:\, iLj} \Phi_j(R, a), \tag{2.19}$$

where Φ_j and U_i are given by (2.8) and (2.9). Let $M_L := \overline{\theta} + (n-1)\lambda$. The baseline assumptions on F continue to apply, and the timing is unchanged.

2.4.3.1 Obedience decisions

Altruistic concerns give a person two reasons to obey: her obedience may reduce her own punishment risk, and it may reduce the punishment risk of someone whose punishment she internalizes. As in the baseline model, obedience takes a cutoff form. A vector x is an equilibrium cutoff profile if and only if, for each i,

$$x_i = \kappa \mathbf{1}\{iRi\} \prod_{\substack{k \neq i \\ iRk}} F(x_k) + \lambda \sum_{j:\, \substack{iLj \\ jRi}} \prod_{\substack{k \neq i \\ jRk}} F(x_k) + \gamma d_i(P). \tag{2.20}$$

The first term is self-pivotality, the ruler punishment that i avoids for herself; the second is altruistic pivotality, punishment imposed on others that i avoids by obeying; and the third is the peer sanction avoided by obedience.

Equation (2.20) also identifies the relevant dependence relation. Person j's obedience raises person i's incentive to obey when it affects the punishment of either i herself or someone whose punishment i internalizes. Equivalently, some person ℓ with $iL\ell$ or $\ell = i$ is responsible for both i and j. Define

$$\widehat{R} := ((L \cup I) \cap R^{-1}) \circ R. \tag{2.21}$$

For $i \neq j$, $i\widehat{R}j$ means that greater obedience by j raises i's gain from obeying. If $L = \varnothing$, then $\widehat{R} = \overline{R}$. The cutoff map remains increasing in others' obedience, so the obedience game again has smallest and largest equilibrium cutoff profiles.

Proposition 2.17 (Obedience with altruistic concerns)**.** *Fix R. Every equilibrium of the obedience subgame is in cutoff strategies, and $x \in [0, M_L]^n$ is an equilibrium cutoff profile if and only if it satisfies* (2.20)*. Moreover, the obedience game has smallest and largest*

equilibrium cutoff profiles, denoted $x^{\min}$ *and* $x^{\max}$*. Both are weakly increasing in* κ*,* γ*, and* λ*, and, with respect to set inclusion, in* L *and* P*.*

The cutoff and lattice structure of the baseline model therefore survives, although the comparative statics with respect to the responsibility relation require more care. An off-diagonal responsibility edge may weaken a person's direct incentive by exposing her to punishment for another person's action; the same edge, however, may strengthen obedience by making her action pivotal for punishment imposed on someone she cares about.

Remark 2.18. Fix P. The comparative statics in Theorem 2.4 with respect to R hold uniformly over responsibility relations if and only if $L = \varnothing$.

The proof constructs both failures: an off-diagonal edge can create altruistic pivotality and raise obedience, while a self-responsibility link can dilute altruistic pivotality and lower it. Altruistic concerns therefore remove the unambiguous effect of responsibility links found in the baseline model. The sign of an added edge depends on the exposure it creates and the altruistic pivotality it generates.

2.4.3.2 Policing decisions

People police others to reduce the expected punishment costs they internalize. For any cutoff vector x, let

$$C_i(x) := \kappa\, \mathbb{E}_{a\sim x}[\Phi_i(R,a)] + \lambda \sum_{k:\, iLk} \mathbb{E}_{a\sim x}[\Phi_k(R,a)]$$

denote the expected punishment cost internalized by i. Policing j is valuable to i only when raising j's obedience lowers $C_i(x)$.

This can occur directly or indirectly: person j's obedience may directly reduce punishment imposed on i or on someone i cares about, or it may raise another person's incentive to obey, which then reduces a punishment cost that i internalizes. The relation $\widehat{R}$ describes these indirect obedience effects. Define

$$\overset{R}{\rightsquigarrow} := (L \cup I) \circ R \circ (I \cup \widehat{R}^{+}). \tag{2.22}$$

Thus $i \overset{R}{\rightsquigarrow} j$ means that greater obedience by j can lower the punishment cost internalized by i. If $L = \varnothing$, then $\widehat{R} = \overline{R}$ and $\overset{R}{\rightsquigarrow} = \overset{R}{\rightarrow}$, recovering the baseline connection relation.

Proposition 2.19 (Policing under altruistic concerns)**.** *Fix $\kappa, \gamma, \lambda > 0$ and the relation L. There exist thresholds $0 < \underline{\xi} \leq \overline{\xi}$ such that, for every responsibility relation R, every on-path policing relation P, and every pair $i \neq j$:*

(i) if iPj, then $i \overset{R}{\rightsquigarrow} j$ and $\xi \leq \overline{\xi}$;

(ii) if $\xi \leq \underline{\xi}$, then iPj if and only if $i \overset{R}{\rightsquigarrow} j$.

Policing continues to follow the links through which another person's obedience reduces punishment costs internalized by the potential policer, but altruistic concern expands those links beyond collective exposure. In particular, collective responsibility is no longer necessary for policing: if $R = I$, then $i \overset{R}{\rightsquigarrow} j$ if and only if iLj for distinct i and j; when $\xi \leq \underline{\xi}$, the policing relation is therefore $P = L$. Even under individual responsibility, people police those whose punishment they internalize. Altruism also changes the ruler's useful responsibility links when policing is too costly to arise.

Remark 2.20 (Ruler's choice without policing). If $\xi > \overline{\xi}$, the ruler can restrict attention to responsibility rules satisfying $R \subseteq I \cup L^{-1}$.

Every useful off-diagonal edge then has the form iRj with jLi: the ruler uses i's punishment to discipline j precisely when j cares about i's punishment. Punishing someone the disobedient person cares about can therefore deter even without policing. In kin-based punishment, this separates two possible channels: relatives may police because the disobedient person's conduct exposes them to punishment, while she may obey because she internalizes the punishment imposed on them.

2.4.4 Noisy ruler monitoring

The baseline assumes that the ruler perfectly observes disobedience. A general treatment of imperfect monitoring would introduce additional strategic problems, including environments

in which both the ruler and the population are uncertain about realized behavior. This section provides a negative answer to a narrower question: does collective responsibility remain useful when the ruler's individual signals are almost uninformative?

Maintain the baseline timing and let the ruler observe a conditionally independent binary signal $s_i \in \{0,1\}$ for each person, where $s_i = 1$ is a bad signal. For some $\alpha \in [0,1)$ and $\delta \in (0, 1-\alpha]$,

$$\Pr(s_i = 1 \mid a_i = 0) = \alpha, \qquad \Pr(s_i = 1 \mid a_i = 1) = \alpha + \delta.$$

The parameter δ measures the informativeness of the signal. Under responsibility relation R, person i is punished if at least one person j for whom iRj receives a bad signal.

Proposition 2.21 (Noisy ruler monitoring)**.** *Fix $\alpha \in [0,1)$ and the remaining primitives of the baseline model. There exists $\overline{\delta} > 0$ such that, for every $\delta \in (0, \overline{\delta})$, individual responsibility is uniquely optimal.*

When δ is small, policing another person hardly changes the policer's probability of ruler punishment: policing may induce greater obedience, but the ruler's signal distribution changes by only δ, so the resulting reduction in punishment risk vanishes as δ approaches zero. The policing cost does not vanish, and sufficiently noisy monitoring therefore eliminates peer enforcement. Because the policing gain is of order δ regardless of the peer sanction, no fixed value of γ restores collective responsibility near $\delta = 0$.

Once policing disappears, responsibility for others has only its direct cost. Any departure from individual responsibility either removes a person's liability for her own signal or makes her punishment depend on additional noisy signals; both changes weaken the connection between her own action and her punishment. Individual responsibility preserves the limited behavioral information contained in the ruler's signals.

2.5 Conclusion

A ruler who knows what each person has done may still punish one person for another's conduct. Holding someone responsible for another's disobedience gives them a reason to police that person, but it also weakens their own incentive to obey: once their punishment depends on what others do, obedience no longer guarantees safety. The ruler must therefore decide where horizontal deterrence is worth the loss of vertical deterrence.

Peer enforcement requires rulers and peers to observe individual disobedience and sanction it. Individual responsibility is uniquely optimal when policing is too costly, when the ruler's sanctions are too weak, or when the ruler's monitoring is too poor. Poor monitoring may still motivate collective punishment through uncertain attribution, but it cannot sustain the mechanism studied here. The same information that permits individual punishment, then, is what can make collective responsibility an effective instrument of peer enforcement.

3 Denunciations under imperfect monitoring

> The people were to be grouped in units of five and ten households, exercising mutual surveillance and mutually responsible before the law. Anyone failing to report an offence was to be cut in two at the waist.
>
> Sima Qian, *The Biography of Lord Shang*

3.1 Punishing silence

Rulers often depend on ordinary people for information about conduct they cannot directly observe. Some regimes do more than solicit or reward such information: they punish people who fail to report disobedience around them. A person may therefore be sanctioned not for what she did, but for what she knew and left undisclosed. This essay studies when punishment for silence increases obedience and how broadly the resulting reporting duties should extend.

I model a denunciation regime as a partition of the population into reporting groups. The ruler independently detects each act of disobedience with some probability, while people observe realized disobedience but do not know what he has detected. If every detected act within a group has been reported, punishment is targeted at those who were named; if some detected disobedience remains unreported, the entire group is punished. Reporting can therefore expose an offender to punishment while protecting the reporter from liability for silence.

This rule creates a basic trade-off. Reports increase the expected cost of disobedience by exposing offenders whom the ruler might otherwise miss, but reporting duties also burden people who obey. Someone who complies may still be punished because another member of her group disobeyed, and avoiding that punishment may require her to file a costly report. Denunciation improves obedience only when the deterrence created by peer reporting outweighs the loss of protection associated with obedience itself.

Historical denunciation regimes often relied on this logic. Shang Yang's reforms in Qin organized households into groups of five and ten whose members were required to monitor one another, report offenses, and bear punishment when offenses went unreported (Ch'ü, 1961). Related duties appeared in institutions such as the Spanish Inquisition and in modern authoritarian regimes (Gellately, 1996; Fitzpatrick and Gellately, 1997; Bergemann, 2019). These systems differed greatly in organization and purpose, but they shared an attempt to extract local information by making silence independently punishable. The mechanism requires *two-sided ignorance*: people may know about conduct the ruler missed, while remaining uncertain about what he already knows.

The next two subsections summarize the results and review the related literature. Section 3.2 presents the model and characterizes reporting, obedience, and the ruler's optimal partition. Section 3.3 introduces imperfect peer monitoring and studies accusations unsupported by direct evidence. Section 3.4 concludes.

3.1.1 Summary of results

The first result isolates the role of two-sided ignorance. If the ruler's detected set is publicly observed, reports provide no new information, and individual punishment uniquely maximizes obedience. The value of denunciation is therefore non-monotonic in the ruler's monitoring capacity. When detection is rare, unreported disobedience is unlikely to be exposed, so punishment for silence generates little reporting. When detection is nearly complete, direct monitoring already provides strong deterrence and reports add little information. Denunci-

ation can improve on direct monitoring only between these extremes, when the ruler detects enough disobedience to make silence dangerous but still benefits from people's information.

The ruler faces a related trade-off over group size. Larger groups create more potential reporters, but they also expose each person to more misconduct by others and make obedience less protective. Even as reporting becomes arbitrarily cheap, optimal reporting groups remain bounded. In the limit, their size is no greater than $\lceil 1/\mu \rceil$ and is weakly decreasing in the ruler's monitoring capacity.

Finally, Section 3.3 allows people to observe peer disobedience imperfectly. When peer monitoring is sufficiently accurate, reports remain confined to observed misconduct, and denunciation can still improve on direct monitoring. When reports become too cheap, however, people may name peers they have not observed disobeying. Such accusations protect reporters from collective punishment but make punishment less connected to actual conduct. With imperfect peer monitoring, denunciation therefore improves obedience only at intermediate reporting costs.

3.1.2 Related literature

This essay contributes to work on denunciation and authoritarian information gathering. Historical studies document how regimes elicited accusations through formal duties, rewards, and social pressure (Gellately, 1996; Fitzpatrick and Gellately, 1997; Bergemann, 2019), while recent empirical work traces how coercive pressure can generate cascades of denunciation (Estévez et al., 2024). The closest formal analysis is Wolton and Yu (2026), who study how an autocrat chooses the anonymity and rewards of denunciations when reports provide noisy information about citizens' loyalty. Their central design problem is how these instruments affect the volume and informativeness of reports, and they also consider anti-regime action. Here disobedience is endogenous from the outset: the ruler chooses reporting groups and makes silence punishable in order to alter incentives to obey. False or baseless accusations arise through different mechanisms. In their model, the chosen reward and anonymity

regime can induce accusations that convey little information about loyalty; here, naming an unobserved peer can protect the reporter from collective punishment.

This essay also relates to theories of collective sanctions and decentralized enforcement. The previous essay studies how collective responsibility induces peer policing when conduct is observable; here liability for silence elicits reports under imperfect state monitoring. Halac et al. (2024) study a principal who partitions agents into monitoring teams and conditions rewards on each team's joint performance in order to uniquely implement effort. Unlike the ruler in this essay, their principal always benefits from finer partitions. More broadly, principals can exploit agents' information about one another through group incentives and delegated monitoring (Varian, 1990; Rahman, 2012). Collective sanctions can induce group members to enforce compliance (Heckathorn, 1990; Levinson, 2003), while local information can sustain enforcement when conduct is better observed within groups (Fearon and Laitin, 1996). In Zhou (2021), rebels trade off group size against secrecy, and the ruler may punish those who knew of a rebellion but failed to report it; here the ruler himself designs reporting groups to maximize obedience. Related work on whistleblowing studies strategic information transmission under organizational incentives and unverifiable reports (Ting, 2008; Chassang and Padró i Miquel, 2019). Here reporting incentives arise from two-sided ignorance and exposure to collective punishment.

3.2 Model of denunciations

3.2.1 Model setup

There are a ruler and a finite set of people $N := \{1, \ldots, n\}$. Each person $i \in N$ chooses whether to obey, $a_i = 0$, or disobey, $a_i = 1$. An action profile is $a = (a_i)_{i \in N} \in \{0, 1\}^n$, and the realized set of disobedient people is $A := \{i \in N : a_i = 1\}$. Each person i has a privately observed gain from disobedience, θ_i, drawn independently from a common distribution F,

which is continuous and strictly increasing on the real line and has full support. The direct payoff from disobedience is $a_i\theta_i$.

The ruler observes disobedience imperfectly. For each person i, he privately receives an independent detection signal $s_i \in \{0,1\}$, with $\Pr(s_i = 1 \mid a_i) = \mu a_i$, where $\mu \in (0,1)$ is his monitoring capacity. Thus $s_i = 1$ means that i has been detected disobeying, while $s_i = 0$ means that no disobedience by i has been detected. Let $S := \{i \in N : s_i = 1\}$ be the detected set.

After actions are taken, people observe A but not S. Each person i then submits a message $M_i \subseteq N$. A message is costly: person i pays $\xi|M_i|$, where $\xi > 0$ is the cost of reporting each name. Let $M := \bigcup_{i \in N} M_i$ be the aggregate message. Whether people have an incentive to denounce each other depends on the ruler's choice of a *denunciation regime*.

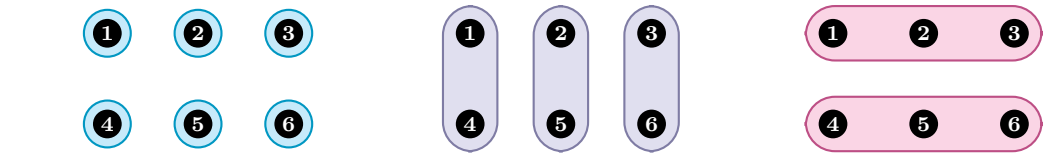
Definition 3.1 (Denunciation regime)**.** A *denunciation regime* is a partition $\mathcal{D}$ of N. Given detected set S and aggregate message M, the punished set $\Pi_{\mathcal{D}}(S, M) \subseteq N$ is defined block by block:

$$\Pi_{\mathcal{D}}(S, M) = \bigcup_{B \in \mathcal{D}} \begin{cases} M \cap B & \text{if } S \cap B \subseteq M, \\ B & \text{otherwise.} \end{cases} \tag{3.1}$$

The rule evaluates coverage at the block level through the aggregate message M: detected disobedience in B is covered if some report names it, so one member's report can protect the whole block from collective punishment.

Figure 3.1 depicts three possible partitions of a population of six people. To illustrate the rule, consider the middle case, where people are divided into pairs. In a pair $B = \{i, j\}$, person i is punished if and only if $i \in S$, or $i \in M$, or $j \in S$ and $j \notin M$. A message naming j can protect i from collective punishment, while a message naming i exposes i to targeted punishment. Messages are therefore accusatory and exculpatory at the same time: they name someone for punishment and help shield the sender from liability for silence.

The timing is shown in Figure 3.2. First, the ruler publicly chooses a partition $\mathcal{D}$. Second, nature draws private disobedience gains $(\theta_i)_{i \in N}$ independently from F. Third, each person



$\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}$ $\{\{1,4\},\{2,5\},\{3,6\}\}$ $\{\{1,2,3\},\{4,5,6\}\}$

(a) Singletons (b) Pairs (c) Triples

Figure 3.1. Denunciation regimes as partitions of the population. People are expected to report disobedience by others in their own block. The ruler's problem reduces to finding the block size that maximizes obedience.

privately observes her own type and chooses whether to obey. Fourth, nature generates detection signals S, which are privately observed by the ruler. Fifth, people simultaneously submit messages without observing S. Finally, punishments are implemented according to $\Pi_{\mathcal{D}}(S, M)$.

The ruler's payoff is $v(a)$, where $v : \{0,1\}^n \to \mathbb{R}$ is strictly decreasing in each coordinate. Each person i's utility is

$$u_i(a, \mathcal{D}, M, S) = \theta_i a_i - \mathbf{1}\{i \in \Pi_{\mathcal{D}}(S, M)\} - \xi|M_i|. \tag{3.2}$$

Punishment is a fixed cost, independent of type, and reporting costs are linear in the number of names reported.

This specification isolates the logic of denunciation under imperfect monitoring. The ruler's signal produces hard evidence: a detected person has disobeyed, while an undetected

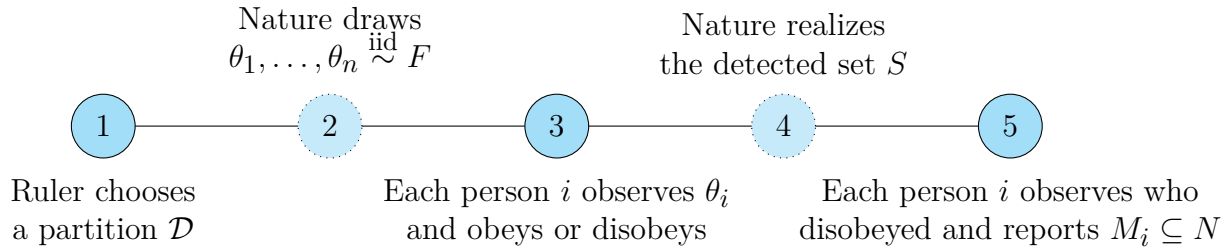


Figure 3.2. Timing of the denunciation model. The ruler observes the detected set S, while people observe the realized disobedience set A but not S when they report.

person has not been cleared. People observe the realized disobedience set, but not the ruler's detections. Messages are useful because they link these two information sources: a message can identify disobedience that the ruler missed, and it can also show that the sender did not remain silent when someone in her block was detected.

The reporting cost ξ captures the friction associated with naming others. It may reflect effort, stigma, fear of retaliation, or the cost of producing a usable statement. Reports are costly, so people must be given an incentive to make them. Denunciation therefore has two effects on obedience incentives: it increases the expected punishment from disobedience, but it can also lower the payoff from obedience by making compliance require costly reporting.

As in the other essays, punishment is bounded and the ruler does not pay a direct cost of punishing. The punishment cost borne by each person is normalized to one. The bound matters: without it, weak monitoring could be offset mechanically by making detected disobedience arbitrarily costly. The absence of a punishment cost favors denunciation, since silence liability expands the set of states in which sanctions are imposed.

A strategy for person i is a pair $\sigma_i := (\alpha_i, \beta_i)$. The obedience rule $\alpha_i : \mathbb{R} \to [0,1]$ gives the probability that i disobeys as a function of her type. The messaging rule $\beta_i : 2^N \to \Delta(2^N)$ gives her message after observing A. The messaging rule need not depend on θ_i: once A is realized, the direct payoff from obedience or disobedience is sunk, so type has no further payoff relevance.

Throughout, equilibrium means perfect Bayesian equilibrium. I restrict attention to symmetric equilibria and to undominated reporting continuations. Because names have no payoff relevance beyond the partition, symmetry means invariance under relabelings that preserve the partition. For $\varphi \in \mathrm{Sym}(N)$ and $X \subseteq N$, write $\varphi(X) := \{\varphi(i) : i \in X\}$. A relabeling preserves $\mathcal{D}$ if $\{\varphi(B) : B \in \mathcal{D}\} = \mathcal{D}$. Let $\mathrm{Sym}(\mathcal{D})$ be the set of relabelings that preserve $\mathcal{D}$.

Definition 3.2 (Symmetric equilibrium)**.** An equilibrium σ is *symmetric* if, for every $\varphi \in \mathrm{Sym}(\mathcal{D})$, $\alpha_{\varphi(i)}(\theta) = \alpha_i(\theta)$ and $\beta_{\varphi(i)}(\varphi(M_i) \mid \varphi(A)) = \beta_i(M_i \mid A)$ for every i, θ, A, and M_i.

This restriction rules out behavior based only on names while allowing strategies to depend on the block structure and on the realized disobedience set.

The analysis proceeds by backward induction. Section 3.2.2 first considers the benchmark in which the detected set is public. Section 3.2.3 solves the reporting problem within a block. Section 3.2.4 derives the obedience cutoff induced by each block size. Section 3.2.5 then studies the ruler's optimal partition.

3.2.2 Public detection

To isolate the role of two-sided ignorance, first consider the benchmark in which the detected set is common knowledge at the reporting stage.

Theorem 3.3 (Public detection)**.** *Suppose S is publicly observed before messages are chosen. Fix any denunciation regime $\mathcal{D}$. In every equilibrium, for every block $B \in \mathcal{D}$, every person $i \in B$, and every message M_i chosen with positive probability, $M_i \subseteq S \cap B \setminus \{i\}$. Under symmetric equilibria, the singleton regime uniquely maximizes the ruler's expected payoff* $\mathbb{E}[v(a)]$.

Public detection removes the informational role of reports. A message outside one's block cannot affect one's punishment, and a message naming an undetected person cannot help cover the detected set. A self-report only exposes the sender to punishment. Hence equilibrium messages can name only detected block members other than oneself. With public detection, larger blocks cannot create informational deterrence. If person i disobeys and is detected, she is punished regardless of the messages she submits. If she obeys or disobeys without being detected, her reporting continuation value is the same. The marginal punishment cost of disobedience therefore comes only from direct detection.

Singleton blocks hence attain the maximal marginal deterrence. A disobedient singleton is punished if detected, while an obedient singleton is never punished and never pays reporting costs. Nonsingleton blocks expose obedient and undetected people to punishment risk generated by others' detected disobedience. Since $F(1) < 1$, every nonsingleton block

faces a positive probability that another member disobeys and is detected. Under symmetry, this strictly lowers obedience relative to the singleton regime. The value of denunciation therefore requires two-sided ignorance: people must know disobedience the ruler may have missed, and they must be uncertain about which disobedients the ruler has detected.

3.2.3 Reporting decisions

Fix a partition $\mathcal{D}$ and a realized disobedience set A. Messages are chosen after people observe A, but before they observe the ruler's detected set S. This informational asymmetry allows messages to reveal disobedience the ruler may have missed.

Lemma 3.4 (Local reports)**.** *Fix a denunciation regime $\mathcal{D}$ and a disobedience set A. In every equilibrium, reports contain only local disobedients other than oneself: for every block $B \in \mathcal{D}$, every person $i \in B$, and every message M_i used with positive probability, $M_i \subseteq A \cap B \setminus \{i\}$.*

Lemma 3.4 removes messages that cannot affect the sender's punishment risk. A message outside one's block has no effect on one's block. A message naming an obedient person cannot help cover the detected set. A self-report only makes the sender more likely to be punished. Hence the continuation problem separates by block.

Fix a block $B \in \mathcal{D}$ of size b, and let $k := |A \cap B|$ be the number of disobedients in the block. For each person $i \in B$, write $M_i^{\max} := A \cap B \setminus \{i\}$ for the maximal local report available to i. The next lemma reduces the reporting problem to two endpoints. Once reports are local, a person either reports all available local disobedients or reports no one. Fix the messages submitted by everyone other than i. Person i's message matters only when some detected disobedients in her block remain uncovered by those other messages. If the set of uncovered names is C, then i prevents collective punishment only by reporting every name in C. Reporting a strict subset of C has no protective value.

Symmetry then reduces the comparison to the number of names reported. Let $|M_i^{\max}| = K$, and suppose i reports r names with $0 < r < K$. Under symmetry, this is equivalent to a uniform random r-subset of $M_i^{\max}$. A lottery that reports all K names with probability

r/K and reports no one otherwise has the same expected reporting cost. It covers any single residual name with the same probability and covers any residual set with more than one name with weakly higher probability. Hence every interior report is weakly dominated by a lottery between silence and the maximal local report.

Lemma 3.5 (Endpoint reports)**.** *Fix a block $B \in \mathcal{D}$ and a disobedience set A. In any undominated symmetric reporting continuation, every message M_i used with positive probability by person $i \in B$ satisfies $M_i = \varnothing$ or $M_i = M_i^{\max}$.*

The reporting problem therefore reduces to whether the maximal local report is worth making. The answer depends on whether anyone in the block obeyed.

If $1 \leq k < b$, an obedient person can report all k disobedients. When her report is pivotal for coverage, it prevents collective punishment if at least one of those k disobedients is detected. The per-name value is

$$v_k(\mu) := \frac{1-(1-\mu)^k}{k}, \qquad 1 \leq k < b. \tag{3.3}$$

In particular, $v_1(\mu) = \mu$. If $k = b$, everyone disobeyed. Person i can report only the other $b-1$ members of the block. Her report can help only if she is not detected, and then prevents collective punishment if at least one other member is detected. The per-name value is

$$w_b(\mu) := \frac{(1-\mu)\left[1-(1-\mu)^{b-1}\right]}{b-1}, \qquad b \geq 2. \tag{3.4}$$

The two values differ because the reporter's own action changes the event in which her report can protect her. An obedient reporter is never directly detected, so her report matters whenever some reported disobedient is detected. A disobedient reporter must also avoid detection herself.

When reporting is strictly profitable and there is a unique person who can make the relevant report, she reports with probability one. When there are several symmetric reporters,

each relevant person mixes between silence and her maximal local report so that the others are indifferent.

Proposition 3.6 (Symmetric reporting)**.** *Fix a block of size b with k disobedients. In any undominated symmetric reporting continuation characterized above:*

(a) if $k = 0$ or $k = b = 1$, everyone is silent;

(b) if $1 \leq k < b$, disobedients are silent, and obedient people report all k disobedients with positive probability if and only if $v_k(\mu) > \xi$;

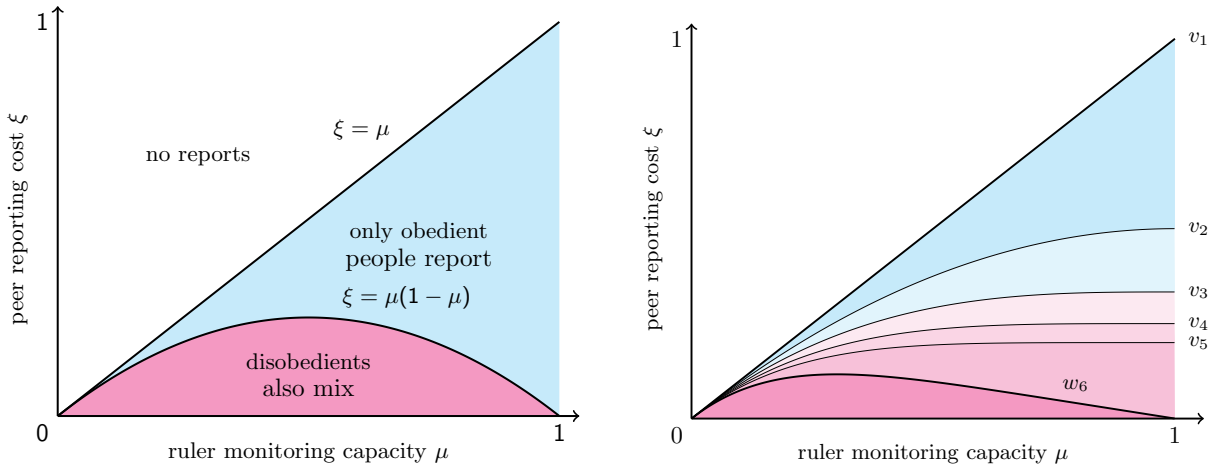
(c) if $k = b \geq 2$, everyone is silent if $w_b(\mu) \leq \xi$, and everyone reports all other block members with positive probability if $w_b(\mu) > \xi$.

Proposition 3.6 gives the reporting continuation for every realized block state. The two-person case is the simplest version: the relevant thresholds are μ and $\mu(1-\mu)$.

Remark 3.7 (Reporting in pairs). Suppose $B = \{i, j\}$. If exactly one person disobeys, say j, then person i can report j. Reporting costs ξ and prevents punishment of i exactly when j is detected, which occurs with probability μ. Hence i reports if and only if $v_1(\mu) = \mu > \xi$.

If both people disobey, then person i can report only j. This report helps i exactly when i is not detected and j is detected, which occurs with probability $\mu(1-\mu)$. Hence disobedients report with positive probability if and only if $w_2(\mu) = \mu(1-\mu) > \xi$.

Figure 3.3 plots these thresholds. Reporting by obedient people becomes easier to sustain as monitoring improves. Reporting in the all-disobedient state is different: a disobedient reporter benefits only when she is missed while someone else is caught, so the threshold is hump-shaped in monitoring capacity.



(a) Block of two

(b) Block of six

Figure 3.3. Symmetric reporting thresholds. Panel (a) shows the three reporting regions in a two-person block. Panel (b) shows the corresponding thresholds in a block of six: $v_k(\mu)$ governs reporting by obedient people when k people disobey, while $w_b(\mu)$ governs reporting in the all-disobedient state.

Proposition 3.6 also pins down the mixing probabilities. For $1 \le k < b$, set

$$\pi^0_{b,k} := \begin{cases} 0, & \text{if } v_k(\mu) \le \xi, \\ 1, & \text{if } v_k(\mu) > \xi \text{ and } b-k=1, \\ 1 - \left(\dfrac{\xi}{v_k(\mu)}\right)^{\frac{1}{b-k-1}}, & \text{if } v_k(\mu) > \xi \text{ and } b-k \ge 2. \end{cases}$$

If $k = b \ge 2$, set

$$\pi^1_b := \begin{cases} 0, & \text{if } w_b(\mu) \le \xi, \\ 1 - \left(\dfrac{\xi}{w_b(\mu)}\right)^{\frac{1}{b-1}}, & \text{if } w_b(\mu) > \xi. \end{cases}$$

Thus the reporting continuation is governed by two thresholds. The function $v_k(\mu)$ applies when at least one person obeyed; reports then come from obedient people and name all disobedients in the block. The function $w_b(\mu)$ applies in the all-disobedient state; reports then come from disobedients and can name only other block members.

3.2.4 Obedience decisions

For a singleton block, the local reporting problem is empty. A person who obeys is never punished, and a person who disobeys is punished exactly when she is detected. Hence the singleton cutoff is $x_1 = \mu$.

Now fix a block B of size $b \geq 2$. Consider a symmetric cutoff strategy in which a person obeys if and only if $\theta_i \leq x$. Fix a tagged person $i \in B$, and let $h \in \{0, \ldots, b-1\}$ be the number of other members of B who disobey. Write $c_h(\mu) := 1 - (1-\mu)^h$ for the probability that at least one of those h people is detected.

If i obeys and h others disobey, her continuation cost is

$$O_b(h) := \min\{\xi h, c_h(\mu)\}. \tag{3.5}$$

If $h = 0$, this cost is zero. If $h \geq 1$ and $v_h(\mu) \leq \xi$, she remains silent and faces collective punishment when at least one of the h disobedients is detected. If $h \geq 1$ and $v_h(\mu) > \xi$, her continuation cost is ξh: with a unique obedient person she reports for sure, while with several obedient people the mixed continuation makes reporting and silence payoff-equivalent. Since $v_h(\mu) = c_h(\mu)/h$, the reporting threshold $v_h(\mu) > \xi$ is equivalent to $\xi h < c_h(\mu)$.

If i disobeys, the block contains $h+1$ disobedients. Her continuation cost is

$$D_b(h) = \begin{cases} 1 - (1 - \pi^0_{b,h+1})^{b-h-1}(1-\mu)^{h+1}, & \text{if } h \leq b-2, \\ 1 - (1-\pi^1_b)^{b-1}(1-\mu)^b, & \text{if } h = b-1. \end{cases} \tag{3.6}$$

In the first case, at least one person obeyed and may report all disobedients. In the second case, everyone disobeyed, so reports can come only from other disobedients. In both cases, $D_b(h)$ is one minus the probability that i escapes punishment when she remains silent against the reports of others.

No additional own-reporting cost appears in (3.6). If a disobedient reporter mixes, reporting and silence give the same continuation payoff, so the continuation value can be evaluated at silence without changing the payoff.

Define $\Delta_b(h) := D_b(h) - O_b(h)$. This is the marginal value of obedience when h other members of the block disobey. Block liability affects obedience only through this difference. It can raise the cost of disobedience by exposing disobedients to reports, and it can raise the cost of obedience by making obedient people report or bear liability for silence.

Proposition 3.8 (Obedience equation)**.** *Fix $b \geq 1$ and the reporting continuation characterized in Proposition 3.6. A symmetric cutoff equilibrium exists. For $b = 1$, $x_1 = \mu$. For $b \geq 2$, every symmetric cutoff equilibrium is a fixed point $x = T_b(x)$, where*

$$T_b(x) = \sum_{h=0}^{b-1} \binom{b-1}{h} (1 - F(x))^h F(x)^{b-1-h} \Delta_b(h). \tag{3.7}$$

The reporting stage determines $\Delta_b(0), \ldots, \Delta_b(b-1)$. The cutoff x determines how likely each value of h is. Higher obedience by others shifts probability toward smaller h, so obedience decisions are strategic complements if $\Delta_b(0) \geq \Delta_b(1) \geq \cdots \geq \Delta_b(b-1)$. This ordering holds in two-person blocks. In larger blocks it can fail because potential reporters may free-ride on one another. When multiple symmetric cutoff equilibria exist, the design problem below evaluates each block size at its largest cutoff equilibrium.

3.2.5 Optimal denunciation regimes

For each block size b, let x_b denote the largest symmetric cutoff equilibrium generated by the reporting continuation characterized above, with $x_1 = \mu$. This is the equilibrium used in the ruler's design problem. If a block size admits multiple symmetric cutoff equilibria, x_b is the ruler-favorable cutoff for that block size. Hence a person assigned to a block of size b disobeys with probability $1 - F(x_b)$.

3.2.5.1 When denunciation helps

The preceding analysis allows any payoff $v(a)$ that is strictly decreasing in disobedience. For the block-size comparison, I specialize the ruler's objective to expected disobedience, so maximizing the ruler's payoff is equivalent to minimizing the expected number of disobedient people. If a partition $\mathcal{D}$ has block sizes $b_1, \ldots, b_m$, then expected disobedience is

$$L(\mathcal{D}) = \sum_{r=1}^{m} b_r[1 - F(x_{b_r})]. \tag{3.8}$$

Thus, under this objective, the partition matters only through its block sizes. A block size with a higher cutoff generates lower expected disobedience per person.

The first question is whether collective liability can improve on direct detection. The singleton partition is the benchmark: each person is punished only when directly detected, and the cutoff is $x_1 = \mu$. A nonsingleton block can improve on this benchmark only if reports raise the marginal cost of disobedience by more than they raise the continuation cost of obedience. The next result shows that this improvement is possible only at intermediate monitoring capacity.

Theorem 3.9 (Non-monotonicity in monitoring capacity)**.** *Suppose $n \geq 2$ and take the ruler's objective to be expected disobedience. The following hold.*

(a) If $\xi \geq \mu$, the singleton partition is uniquely optimal.

(b) If $\xi < \mu(1 - \mu)$, the singleton partition is not optimal.

(c) For every $\xi > 0$, there exist $0 < \underline{\mu} \leq \overline{\mu} < 1$ such that the singleton partition is uniquely optimal whenever $\mu < \underline{\mu}$ or $\mu > \overline{\mu}$.

The result gives the basic non-monotonicity. When monitoring is weak relative to reporting costs, reports do not protect reporters often enough to be induced. Nonsingleton blocks then only expose people to punishment risk generated by others. When monitoring is

very strong, a disobedient reporter is unlikely to avoid direct detection herself, so reports in the all-disobedient state lose value. Between these extremes, reports can make disobedience more costly than direct detection alone.

The two-person block gives the simplest sufficient condition for improving on the singleton partition.

Remark 3.10 (Pairs or singletons?). If $\xi < \mu(1-\mu)$, then a two-person block has cutoff strictly above the singleton cutoff μ. Hence replacing two singleton blocks by one pair strictly lowers expected disobedience.

Equivalently, for fixed $\xi < \frac{1}{4}$, pairs improve on singletons whenever

$$\mu \in \left(\frac{1 - \sqrt{1 - 4\xi}}{2}, \frac{1 + \sqrt{1 - 4\xi}}{2} \right). \tag{3.9}$$

Reports must be cheap enough relative to detection, while monitoring must leave room for disobedient reporters to escape direct detection.

Theorem 3.9 establishes that nonsingleton blocks can improve on direct detection, but it does not say how large useful blocks should be. Larger blocks create more potential reporters, while making each report less pivotal and increasing exposure to others' disobedience. The next result shows that useful blocks remain bounded in size.

3.2.5.2 Cheap reports

I now consider the limit as $\xi \to 0$ in a large-population block-size comparison. By this, I mean the comparison of fixed block sizes $b \geq 2$ when the population is large enough that each such block size is feasible and integer divisibility constraints can be ignored. At the maximal cutoff, the only remaining disobedience comes from types above one. Let $\rho := 1 - F(1)$ be the share of people who disobey even under maximal punishment. Singletons rely only on direct detection, so their cutoff remains below maximal deterrence whenever $\mu < 1$. For every fixed block size $b \geq 2$, cheap reports instead sustain maximal deterrence: $x_b(\xi) \to 1$.

The relevant comparison is therefore among nonsingleton blocks, and is governed by how fast they approach this common limit.

Because every fixed nonsingleton block converges to cutoff one, the limiting design problem compares first-order losses from maximal deterrence. For each $b \geq 2$ and $\mu, \rho \in (0,1)$, define

$$A_b(\mu, \rho) := (b-1)\rho + \frac{(b-1)[\rho(1-\mu)]^{b-1}}{1-(1-\mu)^{b-1}}. \tag{3.10}$$

The first term is the exposure created by irreducible disobedients in the rest of the block. The second term is the residual loss from the all-disobedient reporting problem. In the first-order comparison, $[\rho(1-\mu)]^{b-1}$ is the probability that all other block members are irreducible disobedients and escape direct detection, while $1-(1-\mu)^{b-1}$ is the detection event that makes reports among disobedients valuable.

Theorem 3.11 (Cheap reports)**.** *Let $b^*(\mu, \rho)$ be the smallest minimizer of $A_b(\mu, \rho)$ over integers $b \geq 2$ in the large-population block-size comparison.*

(a) As $\xi \to 0$, the limiting optimal nonsingleton block size is $b^*(\mu, \rho)$, and $b^*(\mu, \rho) \leq \lceil 1/\mu \rceil$.

(b) The limiting solution $b^*(\mu, \rho)$ is weakly decreasing in μ.

The theorem gives a bound on the scale of optimal mutual liability. Even when reports become arbitrarily cheap, optimal blocks remain small. If each offender is detected with probability μ, the limiting optimal block contains no more than about $\frac{1}{\mu}$ people. Larger blocks create more potential reporters, but they also expose each person to more irreducible disobedience in the rest of the block. Past the bound, this exposure dominates.

Cheap denunciation yields bounded liability units whose size is calibrated to monitoring capacity. Better monitoring supports smaller units, because fewer people are needed to make silence dangerous. Worse monitoring requires larger units, but only up to the point at which additional members mainly add exposure to disobedience that cannot be deterred.

3.3 Imperfect peer monitoring

The baseline model assumes that people observe the realized disobedience set before reports are filed. I now allow people to miss peer disobedience with some probability. The ruler still partitions the population into blocks, detects each disobedient person with probability μ, and punishes a whole block if some detected disobedience in that block was left unreported. The new force is accusation without evidence: when peer monitoring is imperfect, cheap reports can induce people to name unobserved peers, making punishment less connected to actual disobedience.

3.3.1 Signals and reporting

Fix a block B of size b, and let $A_B := A \cap B$ be the set of disobedient people in the block. Person $i \in B$ observes a private signal $Z_i \subseteq A_B \setminus \{i\}$: each $j \in A_B \setminus \{i\}$ is independently observed with probability $\nu \in [0, 1]$. People receive no direct signal of obedience. Thus, if $j \notin Z_i$, person i knows only that j either obeyed or disobeyed without being observed.

The parameter ν measures peer-monitoring accuracy. The baseline model is the case $\nu = 1$, in which $Z_i = A_B \setminus \{i\}$ with probability one. When $\nu < 1$, unobserved names are ambiguous: in a symmetric cutoff equilibrium with obedience probability q, the posterior probability that an unobserved person disobeyed is

$$\pi(q, \nu) = \Pr(j \in A_B \mid j \notin Z_i) = \frac{(1-q)(1-\nu)}{q + (1-q)(1-\nu)}. \tag{3.11}$$

This posterior is decreasing in ν, increasing in $1 - q$, equal to zero when $\nu = 1$, and positive when $\nu < 1$ and $q < 1$. This is the source of baseless accusation.

The reporting problem again collapses to endpoints. The signal creates two ordered classes of names: observed disobedients are known to be disobedient, while unobserved names are only possibly disobedient. Thus, in any undominated symmetric continuation, reports first exhaust the observed set before adding unobserved names. Below the signal,

the endpoint argument from Lemma 3.5 applies to the exchangeable set Z_i. Above the signal, the same endpoint argument applies to the exchangeable set $B \setminus (\{i\} \cup Z_i)$.

Thus, in any undominated symmetric reporting continuation, person i's report after signal Z_i can be restricted to $\varnothing$, Z_i, or $B \setminus \{i\}$. These are silence, evidence, and accusation. Silence names no one; evidence reports exactly the observed disobedients; accusation reports every other member of the block, including unobserved names.

Lemma 3.12 (Endpoint reporting with imperfect peer monitoring)**.** *Fix a block B and a signal Z_i. In any undominated symmetric reporting continuation, every report used with positive probability is equivalent to one of three reports: $\varnothing$, Z_i, or $B \setminus \{i\}$. If $\nu = 1$ and $\xi > 0$, accusation is strictly dominated by evidence whenever $Z_i \neq B \setminus \{i\}$.*

The reduction follows from the same convexification argument as in the baseline reporting game. Interior reports of observed disobedients are weakly dominated by lotteries over reporting none and reporting all observed disobedients. Reports that include all observed disobedients and an interior number of unobserved names are weakly dominated by lotteries over adding no unobserved names and adding all unobserved names. Expected reporting costs are unchanged, and the endpoint lottery weakly raises the probability of covering every pivotal detected disobedient person.

When $\nu = 1$, every unobserved person is known to have obeyed, so adding unobserved names cannot cover detected disobedience and only adds cost. When $\nu < 1$, unobserved people may have disobeyed, so accusation can be privately valuable.

Proposition 3.13 (Accurate peer monitoring)**.** *Fix a finite block size b and a reporting cost $\xi > 0$. There exists $\overline{\nu} < 1$ such that, for all $\nu > \overline{\nu}$, accusation is not optimal after any signal with unobserved names. Hence, for sufficiently accurate peer monitoring, all reports are contained in private evidence.*

As $\nu \to 1$, the posterior in (3.11) approaches zero uniformly over equilibrium cutoffs, because finite blocks and bounded reporting costs keep cutoffs in a compact interval and hence

keep obedience probabilities bounded away from zero. The expected benefit from adding unobserved names becomes arbitrarily small, while the reporting cost remains positive.

Proposition 3.14 (Useful denunciation under accurate peer monitoring)**.** *Fix $\mu \in (0,1)$ and suppose $\xi < \mu(1-\mu)$. There exists $\overline{\nu} < 1$ such that, for every $\nu > \overline{\nu}$, a two-person block generates a strictly larger cutoff than singleton punishment. Hence, for sufficiently accurate peer monitoring, pairs strictly improve on singleton punishment throughout the pair-improvement region.*

The proposition follows from continuity around the case of perfect peer monitoring. In the baseline model with $\nu = 1$, pairs strictly improve on singleton punishment by Remark 3.10. By Proposition 3.13, accusation is not optimal when ν is sufficiently close to one, and the two-person reporting continuation converges to the perfect-peer-monitoring continuation.

3.3.2 Design implications

For each block size b, let $x_b(\xi, \nu)$ denote the largest symmetric cutoff generated by an undominated symmetric endpoint reporting continuation of the kind described above. For singleton blocks, reporting plays no role, so $x_1(\xi, \nu) = \mu$. If a partition $\mathcal{D}$ has block sizes $b_1, \ldots, b_m$, optimized expected obedience is

$$Q(\xi, \nu) = \max_{\mathcal{D}} \frac{1}{n} \sum_{B \in \mathcal{D}} |B| F(x_{|B|}(\xi, \nu)). \tag{3.12}$$

The singleton benchmark is $F(\mu)$.

Proposition 3.15 (Cheap-report collapse)**.** *Fix $\nu \in (0,1)$ and a finite population. Then*

$$\limsup_{\xi \downarrow 0} \max_{2 \le b \le n} x_b(\xi, \nu) \le \mu.$$

Equivalently, as reporting costs vanish, no nonsingleton block strictly improves on singleton punishment.

Since $F(1) < 1$, some people disobey even under maximal punishment; since $\nu < 1$, unobserved people may have disobeyed without being observed. As $\xi \downarrow 0$, accusation becomes privately attractive whenever it has a positive chance of mattering. Broad accusation then erodes the deterrence advantage of nonsingleton blocks. At exactly $\xi = 0$, the reporting game may have many weak equilibria because reports are free.

The high-cost endpoint is unchanged. If reports are sufficiently costly, they are not filed. Coarse blocks then add liability without generating information, so singleton punishment weakly dominates every nonsingleton partition. The low-cost endpoint is new: by Proposition 3.15,

$$\lim_{\xi \downarrow 0} Q(\xi, \nu) = F(\mu). \tag{3.13}$$

Thus lowering reporting costs too far can make denunciation less useful. Cheap reports induce accusation, and accusation makes punishment less connected to disobedience.

Corollary 3.16 (Interior reporting costs)**.** *Fix $\nu \in (0, 1)$. Suppose there exist $\hat{\xi} > 0$ and a partition $\hat{\mathcal{D}}$ such that*

$$\frac{1}{n} \sum_{B \in \hat{\mathcal{D}}} |B| F(x_{|B|}(\hat{\xi}, \nu)) > F(\mu).$$

Then useful denunciation occurs only at intermediate reporting costs. In particular, $Q(\hat{\xi}, \nu) > \lim_{\xi \downarrow 0} Q(\xi, \nu) = F(\mu)$, and $Q(\hat{\xi}, \nu)$ also exceeds the value attained when reporting costs are sufficiently high.

The condition is satisfied whenever peer monitoring is sufficiently accurate and reporting costs lie in the pair-improvement region from Proposition 3.14. At very high costs, reports are not filed; at very low costs, accusation makes coarse blocks uninformative. Singleton blocks provide a flat outside option; if a nonsingleton partition is useful at all, it must be between these two extremes.

3.4 Conclusion

This essay studies how a ruler can use people's information about one another to deter disobedience when state monitoring is imperfect. A denunciation regime does this by making silence punishable: if every disobedient detected by the ruler is reported, punishment is targeted at the reported people; if a detected disobedient is left unreported, her whole block is punished. The rule therefore creates reporting incentives by weakening the protection normally provided by obedience, since an obedient person can still be punished when someone in her block disobeys and goes unreported.

The analysis identifies the monitoring conditions under which denunciation improves on singleton punishment. When monitoring is too weak, reports are too costly to induce; when monitoring is very strong, direct detection already disciplines disobedience, and disobedient reporters are unlikely to escape detection. Nonsingleton blocks improve on singleton punishment only between these extremes.

The optimal scale of denunciation is also bounded. Larger blocks create more potential reporters, but they expose each person to more disobedience by others. Even when reports become arbitrarily cheap, optimal blocks remain small: the limiting optimal block size is bounded by monitoring capacity.

The extension to imperfect peer monitoring shows why low reporting costs can become counterproductive. If people may miss peer disobedience, then an unobserved person may have obeyed or may have disobeyed without being seen. Cheap reports can then induce accusation without evidence; in the limit, baseless accusations are so frequent that deterrence collapses. Hence, with imperfect peer monitoring, denunciation improves on singleton punishment only at intermediate reporting costs.

The essay leaves open richer punishment rules. A broader design problem could allow varying punishment severity, leniency for informers, corroboration requirements, penalties for false accusations, or rules that weight direct detection and reports jointly.

4 Collective punishment under anonymity

> Finally, Abraham said, "Please don't get angry, Lord, if I speak just once more. Suppose you find only ten good people there." "For the sake of ten good people," the Lord told him, "I still won't destroy the city."
>
> *Genesis* 18:32, CEV

4.1 The problem of anonymous deterrence

Rulers often govern people whose individual actions they cannot observe. An occupying army may know that an attack came from a village but not who is responsible. A government may observe the opposition's vote share in a precinct, but not personal vote choices. A bureaucracy may know that a production quota was missed but not who shirked. In these settings, punishment must be *anonymous*: it is imposed on the whole group, based only on a collective signal of its members' behavior. The practice is sufficiently important in the law of war that Article 33 of the Fourth Geneva Convention expressly forbids it: "No protected person may be punished for an offence he or she has not personally committed." Yet the strategic logic of anonymous deterrence remains under-theorized.

This essay studies which anonymous collective punishment rules create the strongest incentives for obedience. A ruler faces a finite population of people who privately know their inclination to disobey, namely a type drawn from a known distribution. The ruler observes only the total number who disobey. I show that the obedience-maximizing rule is always a threshold: the group is punished if and only if the disobedience count is at or above a

cutoff. Obedience depends on individual pivotality. A person's behavior matters only when it moves the group across the punishment boundary. For a threshold rule, call the count just below the cutoff its tolerance: the number of disobedient others the rule forgives before one more disobedience triggers punishment. The ruler therefore sets the cutoff where this pivotal event is most likely. In equilibrium, the optimal tolerance coincides with a mode of the induced binomial distribution of others' disobedience. I call this property *self-modality*.

Figure 1.3 in the general introduction illustrates the logic. If the threshold is too low, the pivotal event lies in the lower tail of the induced distribution and is too rare to generate much deterrence. If the threshold is too high, it lies in the upper tail and is again too rare. Deterrence is maximized when the rule's tolerance is placed at a mode of the distribution of others' disobedience.

In the costless exact-count benchmark, this essay offers a theory of frequent repression as a structural feature of optimal deterrence. Because the ruler adjusts the punishment rule to the distribution of types he faces, frequent on-path punishment can arise even in a highly obedient population, and similar punishment rates can occur across populations with different obedience rates. Observed repression is therefore a poor proxy for underlying disloyalty or regime weakness. At the same time, self-modality implies restraint. It is not generally optimal to punish at the first observed disobedience, even when punishment itself is costless: zero tolerance maximizes obedience precisely when zero disobedient others is modal from an individual's perspective.

4.1.1 Summary of results

A ruler maximizes obedience over a finite set of people by publicly committing to an *anonymous* punishment rule: a map from the number of disobedient people to a group punishment level. People privately learn their types, drawn independently from a known distribution F, and choose whether to obey. The ruler observes only the disobedience count.

Fixing a punishment rule and a symmetric conjecture, each person's obedience decision is a cutoff in her type (Lemma 4.1). Deterrence is linear in the increments of the punishment rule, and each increment matters only through the event that the person is pivotal at the corresponding count. Concentrating all deterrence at a single boundary is therefore without loss (Lemma 4.2). Theorem 4.3 characterizes the obedience-maximizing rule: the punishment threshold is *self-modal*, with the rule's tolerance coinciding with a mode of the induced binomial distribution of others' disobedience. The optimal cutoff must be validated by the very behavior it induces. Obedience is higher when types favor obedience and lower in larger groups (Proposition 4.4).

Zero tolerance, where the ruler triggers punishment at the first observed disobedience, is optimal only in groups no larger than a distribution-specific number $\zeta(F)$ (Proposition 4.5). Beyond it, the optimal rule tolerates some disobedience, yet punishes on path with probability at least one quarter, converging to one half as the group grows (Theorem 4.7). Frequent punishment is thus a structural feature of optimal anonymous deterrence, not evidence of disloyalty or weak control.

Section 4.3 considers several extensions. With costly punishment, the problem reduces to choosing an obedience target and paying the least expected on-path punishment that sustains it; least-cost rules are upper thresholds in disobedience counts, and higher costs weakly lower both obedience and expected punishment (Proposition 4.10). With public shocks, obedience and the optimal threshold move together, leaving on-path punishment asymptotically pinned at one half across states (Proposition 4.11). The same self-modality logic extends to symmetric interdependence in obedience payoffs under single crossing and uniform interiority conditions (Proposition 4.14) and to any monotone anonymous monitoring technology, with the threshold now in signal space (Proposition 4.21).

4.1.2 Related literature

The problem belongs to a broader literature on authoritarian control under limited information, including models of dictatorship, preference falsification, and political control (Wintrobe, 1990; Kuran, 1991; Hassan et al., 2022; Egorov and Sonin, 2024). The closest substantive paper is Shadmehr and Bernhardt (2011), who study repression and collective action under imperfect information. In their model, the repression technology is fixed. Here, the ruler chooses the punishment rule, subject to the constraint that punishment can depend only on aggregate disobedience. The object is therefore the design of anonymous collective punishment.

Theories of indiscriminate violence and collective punishment study coercion when individual responsibility is difficult to establish (Kalyvas, 2006; Lyall, 2009; Benmelech et al., 2015; Rozenas, 2020). Rozenas (2020), for example, studies demographic targeting as a way to disrupt cross-group collective action. Here, punishment is anonymous and conditioned only on aggregate disobedience. The ruler does not choose which group to target; he chooses where to place the punishment cutoff within a group.

The results have implications for empirical work on state repression (Davenport, 2007; Hill and Jones, 2014; Ritter, 2014). If frequent collective punishment can arise under optimal deterrence, then observed coercion need not be a reliable proxy for latent disloyalty or weak control. This logic also qualifies designs that use public shocks (such as rainfall, elections, internet outages, focal dates, or censorship capacity) to study mobilization and repression (Ritter and Conrad, 2016; King et al., 2013; Christensen and Garfias, 2018; Truex, 2019; Carter and Carter, 2020; Gohdes, 2020; Garbe, 2024). If such shocks also move the ruler's optimal punishment threshold, observed repression may remain stable or even decrease even as underlying disobedience increases.

On the technical side, the model connects team moral hazard to pivotal participation. Team moral-hazard models study incentives when individual actions are inferred from pooled outcomes (Holmström, 1982; Rasmusen, 1987; Legros and Matthews, 1993; Battaglini, 2006).

Models of pivotal participation, including voting and discrete public-good provision, show how individual incentives depend on the probability of changing a collective outcome (Palfrey and Rosenthal, 1983, 1984; Feddersen and Pesendorfer, 1997, 1998). Recent work on pivotal protest applies this logic to finite-population collective action (Gieczewski and Kocak, 2025). Related models of profiling and deterrence also require policy and behavior to be jointly consistent (Knowles et al., 2001; Persico, 2002; Baliga et al., 2020). Anonymous punishment has a similar threshold logic: one person's action matters when it moves an aggregate count across a cutoff. Because punishment is costless in the baseline, the ruler does not face a budget over sanctions; the design question is which aggregate count should trigger punishment. Self-modality answers this question: punishment is most effective when the boundary is placed where individual pivotality is largest.

4.2 Model of anonymous punishment

This essay studies the obedience-maximizing punishment rule under an *anonymity* constraint: punishment cannot condition sanctions on people's identities.

4.2.1 Model setup

There are a ruler and a finite set of people $N := \{1, \dots, n\}$. Each person $i \in N$ obeys ($a_i = 0$) or disobeys ($a_i = 1$). The ruler observes only the number of people who disobey, denoted $|a| := \sum_{i=1}^{n} a_i$, without observing their identities.

First, the ruler publicly commits to an anonymous punishment rule $p : \{0, \dots, n\} \to [0, 1]$, where $p(r)$ is the group punishment level when exactly r people disobey. Second, nature draws types $\theta_1, \dots, \theta_n$ independently from a distribution F, which is continuous and strictly increasing on the real line and has a finite first moment. The type θ_i is person i's private gain from disobedience. Third, each person privately observes θ_i and chooses $a_i \in \{0, 1\}$. Finally, the ruler observes $|a|$, the prescribed punishment $p(|a|)$ is imposed, and payoffs are realized. The timing is shown in Figure 4.1.

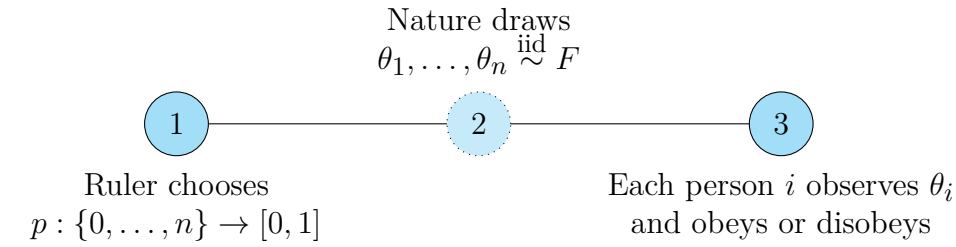


Figure 4.1. Timing of the model of anonymous collective punishment. The ruler's choice is publicly observed, but people learn their types privately. The ruler observes and conditions punishment on the total count of disobedients.

The ruler's payoff is $v(a)$, where $v : \{0,1\}^n \to \mathbb{R}$ is weakly decreasing in each coordinate and non-constant. The ruler's commitment power can be justified by the fact that repression is rarely carried out by the official who designs it: execution is generally delegated to subordinates whose authority is limited to following the rule. Section 4.3.1 considers the case where the ruler pays a cost of punishing, while retaining commitment power.

People's payoffs are additively separable in disobedience gains and punishment: each person i's utility is

$$u_i(a, p) = \theta_i a_i - p(|a|). \tag{4.1}$$

This specification assumes private gains from disobedience do not depend on others' actions. This assumption is adopted for transparency: in the baseline, the only source of strategic interdependence is the punishment rule itself, so pivotality stands out cleanly. Section 4.3.3 relaxes this assumption and shows that the main results hold under arbitrary symmetric action interdependence, including both strategic complements and strategic substitutes, under single crossing in own type and uniform interiority conditions.

A pure strategy for person i is a measurable function $s_i : \mathbb{R} \to \{0,1\}$. Given a punishment rule p, a Bayesian Nash equilibrium is a strategy profile $(s_i)_{i \in N}$ such that each $s_i(\theta_i)$ maximizes expected utility given p and the strategies of the other people. For each rule p, I select the symmetric Bayesian Nash equilibrium with the largest obedience rate. This selection favors the ruler, but it is also natural from people's perspective: as long as obedience imposes no negative externalities on others, and punishment is weakly increasing

in the number of people who disobey, symmetric equilibria with higher obedience are weakly Pareto superior. Remark 4.15 makes this claim precise in the general-payoff case.

4.2.2 Obedience decisions

Fix a punishment rule p. Suppose all people other than i obey with probability q. Since types are independent, the number of other people who disobey is $X_q \sim \text{Bin}(n-1, 1-q)$. If i obeys, total disobedience is X_q; if she disobeys, it is $X_q + 1$. The gain from obedience is

$$-\theta_i + \mathbb{E}[p(X_q + 1) - p(X_q)].$$

Thus i obeys if and only if this gain is non-negative. Since the gain from obedience is strictly decreasing in θ_i, the best response takes a cutoff form.

Lemma 4.1. *For any punishment rule p, there exists a symmetric Bayesian Nash equilibrium in cutoff strategies. In such an equilibrium, there is a threshold t^* such that each person obeys if and only if $\theta_i \leq t^*$.*

Any symmetric cutoff equilibrium is summarized by an obedience rate $q = F(t^*)$. Define the deterrence created by a punishment rule at conjectured obedience rate q by

$$D_p(q) := \mathbb{E}[p(X_q + 1) - p(X_q)], \tag{4.2}$$

where $X_q \sim \text{Bin}(n-1, 1-q)$. This is the expected reduction in punishment from obedience, or equivalently the expected increase in punishment from disobedience. A marginal type is indifferent when $\theta_i = D_p(q)$, so any symmetric equilibrium obedience rate satisfies

$$q = F\Big(D_p(q)\Big). \tag{4.3}$$

Let $q(p)$ denote the largest solution. Since the null rule $p \equiv 0$ sustains the voluntary obedience rate $q^0 := F(0)$, the ruler never needs to consider selected equilibria below q^0.

4.2.3 Self-modal thresholds

Under the largest symmetric equilibrium selection, each punishment rule p induces a product distribution of actions with common obedience probability $q(p)$. Since the ruler's payoff is decreasing in each coordinate, higher selected obedience raises the expected ruler payoff. Hence the ruler can, without loss, choose p to maximize the largest symmetric obedience rate it sustains:

$$\max_{p:\{0,\dots,n\}\to[0,1]} q(p), \tag{4.4}$$

where $q(p)$ is the largest solution to (4.3).

For $r \in \{0,\dots,n-1\}$, let $\Lambda_q(r) := \Pr[\mathrm{Bin}(n-1,1-q)=r]$ denote the probability that exactly r other people disobey. Then

$$D_p(q) = \sum_{r=0}^{n-1} \Lambda_q(r)\Big(p(r+1)-p(r)\Big). \tag{4.5}$$

Deterrence is linear in the increments of the punishment rule, and each increment matters only through the event that exactly r other people disobey. The following lemma shows that the ruler loses nothing by concentrating deterrence at a single boundary. Equivalently, the ruler places the punishment boundary at the count where a person is most likely to be pivotal.

For $h \in \{1,\dots,n\}$, let $p_h(r) := \mathbf{1}\{r \ge h\}$ and define

$$\mathrm{Piv}_h(q) := \Pr\{\mathrm{Bin}(n-1,1-q)=h-1\}.$$

This is the probability that exactly $h-1$ other people disobey, so that one person's disobedience moves the group from safety to punishment. The number $h-1$ is the rule's tolerance.

Lemma 4.2. *For every punishment rule p, there exists a threshold rule p_h such that $q(p_h) \ge q(p)$.*

The ruler therefore loses nothing by using an increasing threshold rule, where the group is punished if the disobedience count is at or above a cutoff and spared otherwise. Under p_h, a person's action matters only when exactly $h-1$ other people disobey. In that event, and only in that event, obedience moves the group from punishment to safety, so $D_{p_h}(q) = \mathrm{Piv}_h(q)$. For a fixed obedience rate q, the best threshold maximizes $\mathrm{Piv}_h(q)$, placing the rule's tolerance at a mode of the distribution of other people's disobedience.

Write $D^*(q) := \max_{1 \le h \le n} \mathrm{Piv}_h(q)$, and define

$$T(q) := F\Big(D^*(q)\Big). \tag{4.6}$$

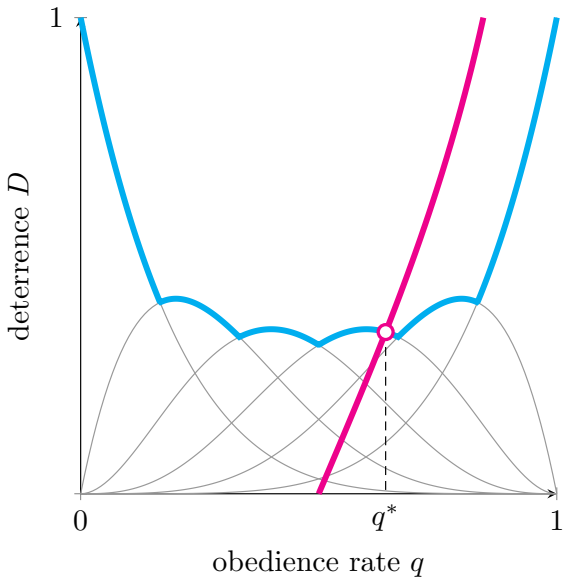
Here $D^*(q)$ is the largest deterrence attainable at obedience rate q, and $T(q)$ is the largest obedience rate sustainable when others obey with probability q and the ruler uses the best threshold. Since each function $\mathrm{Piv}_h(q)$ is continuous in q and the maximum is taken over finitely many h, both D^* and T are continuous. Maximal symmetric equilibrium obedience q^* is then the largest fixed point of T.

Theorem 4.3 (Self-modality)**.** *Let q^* denote maximal symmetric equilibrium obedience. A threshold rule p_h maximizes obedience if and only if $h-1$ is a mode of* $\mathrm{Bin}(n-1, 1-q^*)$. *In particular, the strictest obedience-maximizing disobedience threshold is $h^* = n - \lfloor nq^* \rfloor$.*

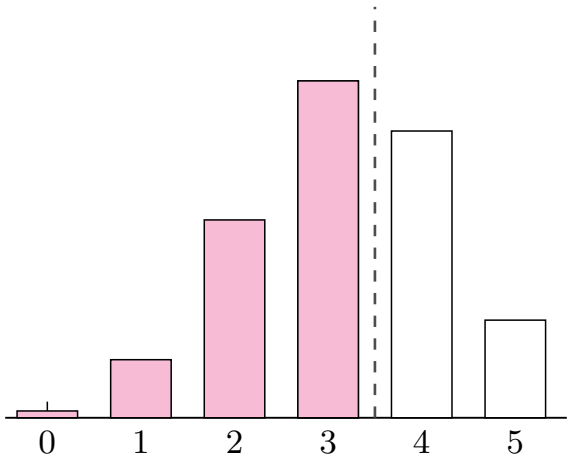
The optimal punishment threshold must be validated by the distribution of other people's disobedience that it induces. In equilibrium, the rule's tolerance lies at a mode of the induced binomial distribution.[1] Equivalently, in obedience-count notation, the largest optimal obedience cutoff is $\lfloor nq^* \rfloor$.

Figure 4.2 illustrates the fixed-point logic behind the optimal threshold. Panel (a) plots the maximal attainable deterrence $D^*(q)$ together with the deterrence required to sustain obedience rate q, namely $F^{-1}(q)$. Their largest intersection determines maximal symmetric

1. If $n(1-q)$ is an integer, $\mathrm{Bin}(n-1, 1-q)$ has two adjacent modes, $n(1-q)-1$ and $n(1-q)$. Selecting the strictest optimal threshold corresponds to selecting the smaller of these two pivotal counts.



(a) Fixed point characterization



(b) Self-modality of optimal punishment

Figure 4.2. Self-modality as a fixed point. Panel (a) plots maximal attainable deterrence $D^*(q)$ in blue and required deterrence $F^{-1}(q)$ in pink; their intersection determines q^*. Panel (b) shows the induced distribution of other people's obedience at q^*. In obedience-count notation, the optimal boundary is placed at a modal count, corresponding to the envelope segment containing the intersection in panel (a). Here $n = 6$ and F is the standard normal, so $q^* \approx 0.63$ and $h^* = 3$.

equilibrium obedience q^*. The function $D^*(q)$ is the largest pivotal mass of $\text{Bin}(n-1, 1-q)$: it is high near $q = 0$ and $q = 1$, where the distribution is concentrated on a few counts, and lower at intermediate values of q, where probability mass is more dispersed. The scalloped shape comes from taking the upper envelope of the curves $\text{Piv}_h(q)$, viewed as functions of q; each scallop is the region of obedience rates where a given cutoff is best. Panel (b) shows the induced distribution of other people's obedience at q^*, with the counts at or below the optimal obedience boundary shown in pink. In obedience-count notation, the largest optimal obedience cutoff is $\lfloor nq^* \rfloor = 3$. Equivalently, the disobedience threshold is $h^* = 3$, so the rule tolerates $h^* - 1 = 2$ disobedient others. Under p_3, the group is punished when total disobedience is at least 3; a person is pivotal exactly when three other people obey.

Proposition 4.4 (Comparative statics of maximal obedience)**.** *Let $q^*(F, n)$ denote maximal equilibrium obedience under distribution F and population size n.*

(a) If $F_L(x) \geq F_H(x)$ for all x, then $q^*(F_L, n) \geq q^*(F_H, n)$.

(b) For every F and n, $q^*(F, n+1) \leq q^*(F, n)$.

Part (a) follows because a gain distribution with more mass below every cutoff shifts the fixed-point map $T(q) = F(D^*(q))$ upward. Part (b) follows because larger populations shift the map downward: any one individual is less likely to be pivotal.

4.2.4 Zero tolerance or frequent punishment

Theorem 4.3 also clarifies the role of restraint. In the obedience-maximization benchmark, punishment is free for the ruler. One might therefore expect zero tolerance, namely $p_1(r) := \mathbf{1}\{r \geq 1\}$: punish the group whenever anyone disobeys. This intuition is correct for a single person, where the ruler uses cutoff $h^* = 1$, and the obedience rate is $F(1)$. For larger groups, however, the ruler maximizes obedience by placing the punishment boundary where one person's action is most likely to be pivotal. Zero tolerance is optimal only when the all-

obedient margin is itself modal; once that fails, the optimal rule tolerates some disobedience before punishment is imposed.

However, this restraint does not make punishment rare on path. Define zero tolerance as the threshold rule $p_1(r) := \mathbf{1}\{r \geq 1\}$. For each $n \geq 1$, let $\widehat{q}_n$ denote the largest equilibrium obedience probability induced by zero tolerance. Under zero tolerance, a person is pivotal exactly when no other person disobeys, so the cutoff equation is $t = \widehat{q}_n^{n-1}$ and $\widehat{q}_n = F(t)$. Equivalently, if x_n is the largest solution of $x = F(x)^{n-1}$, then $\widehat{q}_n = F(x_n)$.

Zero tolerance is obedience-maximizing if and only if zero disobedient others is a mode of $\mathrm{Bin}(n-1, 1-\widehat{q}_n)$. This condition is equivalent to $\widehat{q}_n \geq \frac{n-1}{n}$. Hence zero tolerance is optimal exactly when the expected number of disobedient people in the zero-tolerance equilibrium is at most one.

Define the *zero-tolerance number* by

$$\zeta(F) := \max\left\{n \geq 1 : \widehat{q}_n \geq 1 - \frac{1}{n}\right\}. \tag{4.7}$$

The condition defining $\zeta(F)$ is monotone in n, and it always holds at $n = 1$.

Proposition 4.5 (Zero tolerance in small groups)**.** *Zero tolerance is obedience-maximizing if and only if $n \leq \zeta(F)$.*

The zero-tolerance number is the largest population size for which zero tolerance remains obedience-maximizing. Whenever F is continuous and has full support, $\zeta(F)$ is finite and at least one: zero tolerance is always obedience-maximizing for one person, and there is always a group size large enough that zero tolerance is not obedience-maximizing. At or below this number, punishment can be rare. Under zero tolerance, punishment occurs on the equilibrium path with probability $1 - \widehat{q}_n^n$, which can be small only if the rule induces near-universal obedience.

For any conjectured obedience rate q, let $B \sim \mathrm{Bin}(n, 1-q)$ denote total disobedience and define $h_n(q) := n - \lfloor nq \rfloor$. Under the self-modal threshold, punishment occurs with probability

$P_n(q) := \Pr\{B \geq h_n(q)\}$, equivalently $P_n(q) = \Pr\{O \leq \lfloor nq \rfloor\}$ for $O \sim \text{Bin}(n, q)$. Figure 4.3 shows why this obedience–punishment mapping is discontinuous and non-monotone. On each segment, the threshold is fixed, so punishment declines smoothly with obedience. At the dotted lines, the self-modal boundary changes and punishment jumps. The rightmost branch is zero tolerance, where punishment tracks the probability that at least one person disobeys. The hollow marker in each panel shows the zero-tolerance equilibrium $\widehat{q}_n$ for F the standard normal. The figure shows that zero tolerance passes the self-modality test for $n = 1, 2, 3$ but not 4; hence the standard normal has $\zeta(F) = 3$. The projected point in panel (b) marks the limiting value $\lim_{q \uparrow \frac{1}{2}} P_2(q) = \frac{1}{4}$, which is the sharp infimum in Pinelis (2021). Away from the rightmost branch, punishment is applied near the center of the induced count distribution, so it remains bounded away from zero and quickly settles near one half.

When n exceeds the zero-tolerance number, the obedience-maximizing rule moves the threshold into the body of the induced distribution. Let q_n^* denote maximal equilibrium obedience at population size n, let $h_n^* := n - \lfloor nq_n^* \rfloor$ be the associated disobedience threshold, and let $B_n \sim \text{Bin}(n, 1 - q_n^*)$ be total disobedience. Equilibrium-path punishment occurs with probability $P_n^* := \Pr(B_n \geq h_n^*)$.

Lemma 4.6 (Binomial lower tail)**.** *For $O \sim \text{Bin}(n, q)$, if $\lfloor nq \rfloor \leq n-2$, then $\Pr(O \leq \lfloor nq \rfloor) \geq \frac{1}{4}$. If (q_n) is bounded away from zero and one and $O_n \sim \text{Bin}(n, q_n)$, then $\lim_{n\to\infty} \Pr(O_n \leq \lfloor nq_n \rfloor) = \frac{1}{2}$.*

The first bound is sharp as an infimum: for $n = 2$, $\Pr(O \leq \lfloor 2q \rfloor) = (1-q)^2$ when $q < \frac{1}{2}$, converging to one quarter as q approaches one half (Pinelis, 2021). The second claim follows from the central limit theorem, since $\lfloor nq_n \rfloor$ differs from the mean nq_n by less than one.

Theorem 4.7 (Zero tolerance or frequent punishment)**.** *For every F and every n, either $n \leq \zeta(F)$ or $P_n^* \geq \frac{1}{4}$. Moreover, $\lim_{n\to\infty} P_n^* = \frac{1}{2}$.*

Hence rare punishment is confined to the zero-tolerance region. Once $n > \zeta(F)$, the optimal rule is restrained, in that it allows some disobedience before punishment is imposed.

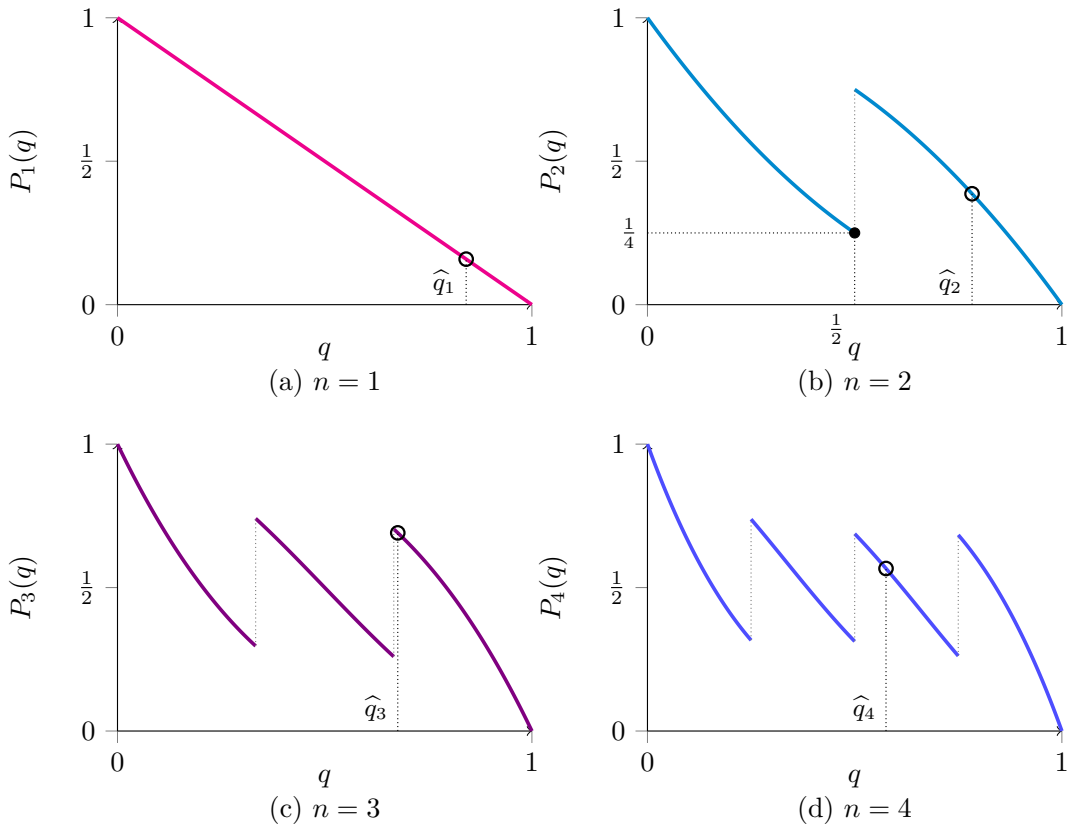


Figure 4.3. Punishment under the self-modal threshold as a function of obedience q and group size n. Each panel plots $P_n(q) = \Pr\{B \geq h_n(q)\}$, where $B \sim \text{Bin}(n, 1-q)$ and $h_n(q) = n - \lfloor nq \rfloor$. Smooth segments correspond to fixed thresholds; jumps occur when the self-modal boundary changes. The rightmost branch is zero tolerance. Hollow markers show the zero-tolerance equilibrium $\widehat{q}_n$ for $F = \mathcal{N}(0, 1)$.

Yet punishment is frequent on the equilibrium path: it occurs with probability at least one quarter, and this probability converges to one half. The reason is that the optimal rule punishes realized disobedience only once it exceeds its typical level, and a random count exceeds its typical level about half the time in large groups.

The same logic explains why obedience provides little individual protection against punishment in large groups. Under the optimal rule, the punishment risks conditional on obedience and disobedience differ exactly by the pivotal mass $D^*(q_n^*)$, which vanishes with n. The rule deters because the threshold is placed where pivotality is as large as possible, even though each person is rarely pivotal.

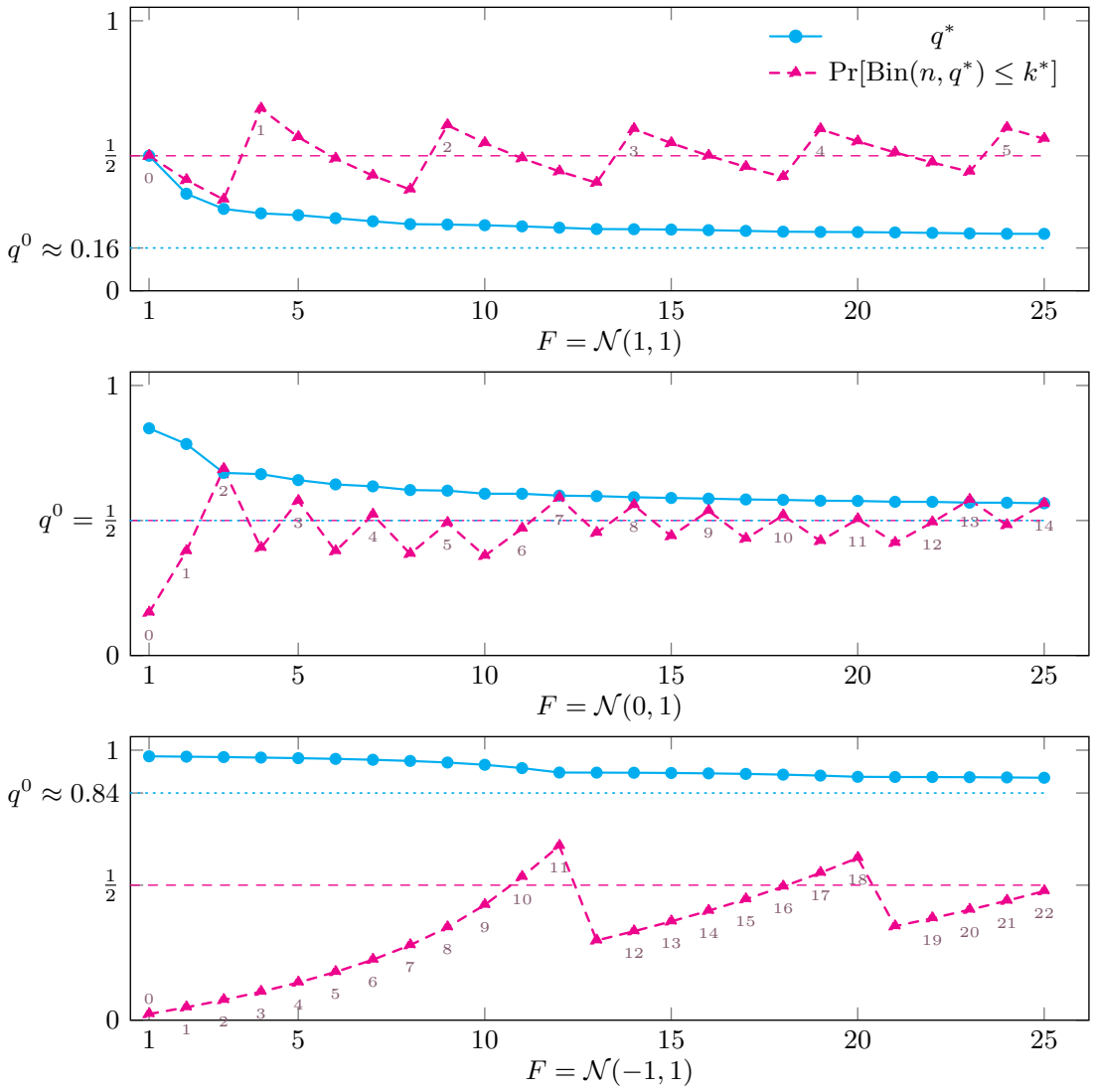


Figure 4.4. Obedience and punishment as group size increases. The horizontal axis indicates population size n. Blue circles plot maximal equilibrium obedience q_n^*, and pink triangles plot equilibrium-path punishment $P_n^* = \Pr(B_n \geq h_n^*)$, where $B_n \sim \mathrm{Bin}(n, 1 - q_n^*)$ and $h_n^* = n - \lfloor nq_n^* \rfloor$. The horizontal blue line marks the voluntary obedience level $q^0 := F(0)$, and the numbers by the pink triangles indicate the obedience cutoff $k_n^* = \lfloor nq_n^* \rfloor$.

Figure 4.4 shows equilibrium obedience and punishment, for three different distributions of types, as n grows. The figure suggests that convergence to one half is fast. In all three examples, on-path punishment is already close to one half for moderate group sizes, even though equilibrium obedience remains visibly above the voluntary level q^0.

4.3 Robustness checks and extensions

4.3.1 Costly punishment

The baseline treats punishment as free. This extension gives the ruler a direct linear cost of punishing. For simplicity, assume that the ruler's utility is now

$$v_\kappa(a,p) = -\frac{|a|}{n} - \kappa p(|a|),$$

where $\kappa \geq 0$ is the cost of punishing. Hence, up to the additive constant -1, if a rule induces selected obedience q and equilibrium-path expected punishment P, his expected payoff is $q - \kappa P$.

At a fixed conjectured obedience rate q, let $X_q \sim \text{Bin}(n-1, 1-q)$ be the number of other people who disobey and $B_q \sim \text{Bin}(n, 1-q)$ the total number who disobey. A rule p generates deterrence $D_p(q)$ and expected on-path punishment $P_p(q)$, and the feasible punishment solid is

$$\begin{aligned} D_p(q) &:= \mathbb{E}[p(X_q+1) - p(X_q)], \\ P_p(q) &:= \mathbb{E}[p(B_q)], \\ \mathcal{B}_n &:= \left\{ (q, D_p(q), P_p(q)) : \begin{array}{l} q \in [0,1], \\ p : \{0,\dots,n\} \to [0,1] \end{array} \right\}. \end{aligned}$$

The floor projection gives the deterrence levels attainable at each obedience rate. Its upper envelope is the deterrence envelope from the previous section. Once punishment is costly, rules that deliver the same (q, D) are no longer equivalent. Among those rules, the ruler chooses the one with the lowest expected on-path punishment.

Figure 4.5 shows this reduction. The pink curve is the deterrence needed to sustain each obedience rate. The frontier turns that deterrence requirement into a punishment cost: sustaining q costs $C(q)$. The orange curve is the ruler's menu after this reduction. He chooses from that menu to maximize $q - \kappa C(q)$.

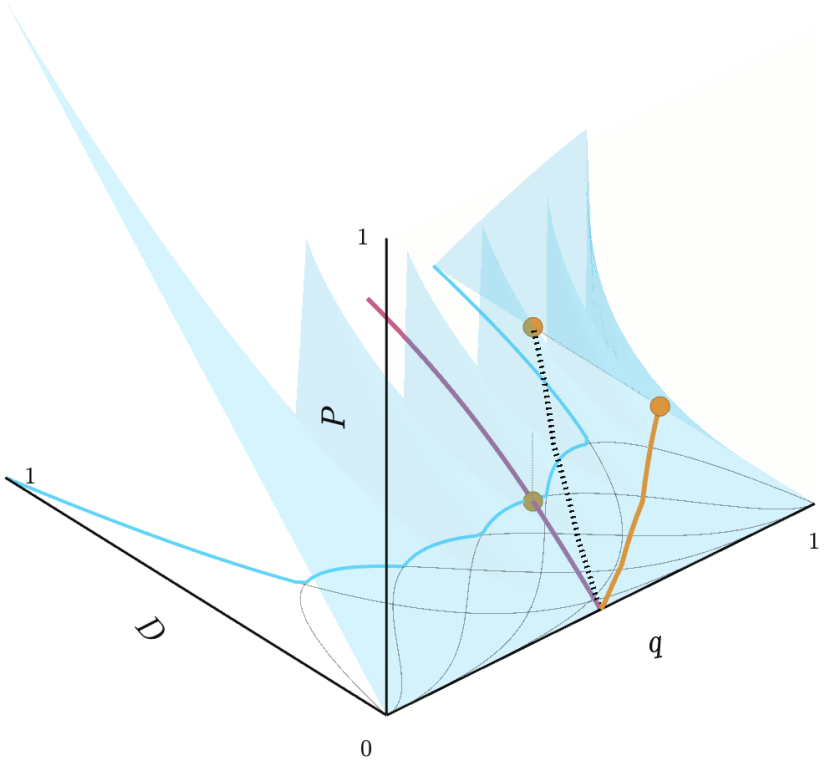


Figure 4.5. The blue surface represents the ruler's frontier, namely the lowest expected punishment P he must give to generate deterrence D at obedience rate q. The pink curve on the floor is the equilibrium deterrence requirement $D = \tau(q)$. The dashed black line is the set of triples $(q, \tau(q), C(q))$. Projecting this set into the (q, P) plane yields the orange curve $P = C(q)$, the ruler's menu. Here $n = 6$ and $F = \mathcal{N}(0, 1)$.

Fix $q \in (0, 1)$. Write $\Lambda_q(r) := \Pr(X_q = r)$, with $\Lambda_q(-1) = \Lambda_q(n) := 0$, and $\beta_q(r) := \Pr(B_q = r)$. Then

$$D_p(q) = \sum_{r=0}^{n} c_q(r)p(r), \qquad c_q(r) := \Lambda_q(r-1) - \Lambda_q(r),$$
$$P_p(q) = \sum_{r=0}^{n} \beta_q(r)p(r).$$

For fixed q, finding the lower surface is a knapsack problem: punishing count r costs $\beta_q(r)$ on path and contributes $c_q(r)$ to deterrence.

For fixed q, define

$$C_q(D) := \min_{0 \le p(r) \le 1} P_p(q) \quad \text{subject to} \quad D_p(q) \ge D,$$

with $C_q(D) = +\infty$ if the constraint is infeasible. Since both $D_p(q)$ and $P_p(q)$ are linear in p, C_q is convex. Whenever $\beta_q(r) > 0$,

$$\frac{c_q(r)}{\beta_q(r)} = \frac{r - n(1-q)}{nq(1-q)}.$$

This ratio is strictly increasing in r. Deterrence is cheapest at the highest disobedience counts. The ruler therefore punishes from the top of the disobedience-count distribution downward, with possible fractional punishment at one boundary count. Equivalently, the cheapest rule may mix between two adjacent upper-threshold rules.

Lemma 4.8 (Punishment frontier)**.** *Fix $q \in (0,1)$ and a deterrence target D. If $C_q(D) < +\infty$, then a least-cost rule can be chosen in upper-threshold form:*

$$p(r) = \begin{cases} 0 & \text{if } r < h, \\ \alpha & \text{if } r = h, \\ 1 & \text{if } r > h, \end{cases}$$

for some $h \in \{1, \dots, n\}$ and $\alpha \in [0,1]$.

The lemma identifies the lower surface in Figure 4.5. For any fixed obedience rate and deterrence target, the cheapest rule starts at the highest disobedience counts and lowers the punishment boundary only as needed. When the required deterrence falls between two pure thresholds, the rule uses partial punishment at the boundary count.

We can now impose equilibrium. In the private-values benchmark, sustaining obedience rate q requires deterrence $\tau(q) := F^{-1}(q)$. Since the null rule already sustains $q^0 := F(0)$ and deterrence is capped by 1, it is enough to consider $q \in [q^0, F(1)]$. Define $C : [0,1] \to [0, +\infty]$

by setting $C(q) = 0$ for $q \leq q^0$, $C(q) := C_q(\tau(q))$ for $q \in (q^0, F(1)]$, and $C(q) = +\infty$ for $q > F(1)$, with $C(q) = +\infty$ if the constraint is infeasible. This is the least expected on-path punishment needed to sustain obedience rate q.

Lemma 4.9 (Target-cost reduction)**.** *Under the maintained highest-obedience selection, the costly-punishment problem is value-equivalent to*

$$\sup_{q:C(q)<+\infty} \{q - \kappa C(q)\}.$$

For each κ, choose an optimal upper-threshold rule, allowing partial punishment at one boundary count. Let q_κ denote its selected equilibrium obedience, P_κ its equilibrium-path expected punishment, and h_κ the lowest disobedience count punished with positive probability. Among optimal upper-threshold rules, choose the largest such boundary, and set $h_\kappa := n + 1$ when the null rule is optimal.

Proposition 4.10 (Costly punishment)**.** *As κ increases, equilibrium obedience and expected on-path punishment weakly decrease, while the punishment boundary weakly increases. The ruler's value weakly decreases. In the private-values benchmark, individuals' ex ante welfare weakly increases.*

A higher punishment cost shifts the ruler toward obedience targets that require weakly less expected on-path punishment. Because least-cost rules punish from the highest disobedience counts downward, the chosen boundary moves weakly upward, and selected obedience falls weakly as the ruler buys less deterrence.

The ruler's value falls because the old rule remains feasible when punishment becomes more expensive. The welfare statement uses the private-values benchmark to isolate the repression effect. A higher punishment cost lowers punishment and removes only obedience that was privately induced by punishment. With interdependent payoffs, lost obedience may create externalities, so individuals' welfare need not have a fixed sign.

Costly punishment limits how far the ruler pushes deterrence. He chooses how much obedience to buy, and the price is the punishment needed to sustain it. Frequent punishment can still occur, but only when its obedience gain justifies its cost.

4.3.2 Public correlation

A public shock can move obedience and the optimal punishment threshold together, leaving on-path repression nearly unchanged. Let $z \in Z$ be a public state, observed by the ruler and all people before the ruler chooses a punishment rule. Conditional on z, types are independently drawn from F_z, and a state-contingent rule is $p_z : \{0, \dots, n\} \to [0, 1]$, where $p_z(r)$ is punishment when exactly r people disobey. State by state, the ruler faces the benchmark problem with F replaced by F_z. The baseline self-modality argument therefore applies state by state: an obedience-maximizing rule can be chosen as a threshold in disobedience counts.

Public shocks move behavior, but they also move the optimal repression. For each population size n and public state z, let $q_{n,z}$ be the selected obedience rate in state z. Under the associated state-contingent threshold rule,

$$h_{n,z} := n - \lfloor nq_{n,z} \rfloor, \qquad P_{n,z} := \Pr\{\mathrm{Bin}(n, 1 - q_{n,z}) \geq h_{n,z}\}.$$

Equivalently, if $O \sim \mathrm{Bin}(n, q_{n,z})$ is the total number of obedient people, then $P_{n,z} = \Pr\{O \leq \lfloor nq_{n,z} \rfloor\}$.

Proposition 4.11 (Public shocks and on-path punishment)**.** *Suppose there exists $\varepsilon > 0$ such that $q_{n,z} \in [\varepsilon, 1 - \varepsilon]$ for every n and every public state z. Under the state-contingent obedience-maximizing threshold rule,*

$$\sup_{z \in Z} \left| P_{n,z} - \frac{1}{2} \right| \to 0.$$

The on-path probability of collective punishment is asymptotically pinned at $\frac{1}{2}$, uniformly across public states. Public shocks can move obedience, thresholds, and realized counts while leaving the equilibrium frequency of collective punishment nearly unchanged.

4.3.3 Interdependent payoffs

The baseline assumes that people's payoff from obedience depends only on their own action and type. I now allow the gain from obedience to depend on others' actions, including through strategic complements or substitutes. Let each person's utility be given by

$$u_i(a, p, \theta_i) = w_i(a, \theta_i) - p(|a|), \tag{4.8}$$

where $w_i : \{0,1\}^n \times \mathbb{R} \to \mathbb{R}$ may depend on the full action profile. The main result survives this generalization given the following assumptions.

Assumption 4.12. The environment satisfies the following conditions.

(a) For every $i \in N$ and every $a_{-i} \in \{0,1\}^{n-1}$, the obedience gain

$$g_i(a_{-i}, \theta) := w_i(0, a_{-i}, \theta) - w_i(1, a_{-i}, \theta)$$

is continuous and strictly decreasing in θ.

(b) For any $i, j \in N$, there exists a permutation $\sigma : N \to N$ with $\sigma(i) = j$ such that $w_j(a_\sigma, \theta) = w_i(a, \theta)$ for all $a \in \{0,1\}^n$ and $\theta \in \mathbb{R}$.

(c) There are $\underline{\theta} < \overline{\theta}$, independent of n, i, and a_{-i}, such that $g_i(a_{-i}, \underline{\theta}) > 1$ and $g_i(a_{-i}, \overline{\theta}) < -1$.

(d) F is continuous on $\mathbb{R}$ and strictly increasing on $[\underline{\theta}, \overline{\theta}]$, with $F(\underline{\theta}) > 0$ and $F(\overline{\theta}) < 1$.

Parts (a) and (c) jointly ensure a unique interior threshold; (b) ensures the expected obedience gain is symmetric across people; (d) keeps the induced obedience rate interior.

Under a symmetric conjecture that others obey independently with probability q, the obedience gain is random because it depends on others' actions. Let

$$G_n(q, \theta) := \mathbb{E}\Big[g_i(A^q_{-i}, \theta)\Big],$$

where each coordinate of A^q_{-i} equals 0 with probability q and equals 1 otherwise. By Assumption 4.12(b), this does not depend on the identity of i. For any $q \in [0, 1]$ and $D \in [-1, 1]$, let $t_n(q, D)$ denote the unique solution to $G_n(q, t) + D = 0$.

Lemma 4.13 (Symmetric reduction)**.** *Fix a conjectured symmetric obedience rate q and a punishment rule p. A person of type θ_i has total gain from obedience $G_n(q, \theta_i) + D_p(q)$. Hence the best response is cutoff at $t_n(q, D_p(q))$, and any symmetric equilibrium obedience rate satisfies*

$$q = F\Big(t_n(q, D_p(q))\Big).$$

Lemma 4.13 is the key formal result of the extension. The baseline analysis has two steps: identifying the marginal type for a given deterrence level, and choosing the punishment rule that maximizes deterrence at a fixed conjectured obedience rate. Interdependence changes only the first step, through G_n: the marginal type is now $t_n(q, D_p(q))$ rather than $D_p(q)$. The second step is unchanged because deterrence retains the binomial pivotality expression in (4.5). The baseline self-modality argument therefore goes through.

Let q^*_n denote maximal equilibrium obedience, $h^*_n := n - \lfloor nq^*_n \rfloor$ the associated disobedience threshold, and $P^*_n := \Pr\{\text{Bin}(n, 1 - q^*_n) \geq h^*_n\}$ equilibrium-path punishment under the largest obedience-maximizing threshold.

Proposition 4.14 (Strategic interdependence)**.** *Under Assumption 4.12, self-modality holds unchanged. If zero tolerance is not obedience-maximizing, then $P^*_n \geq \frac{1}{4}$. Moreover, $P^*_n \to \frac{1}{2}$ as $n \to \infty$.*

Zero tolerance is still characterized by the self-modality test. Let $\hat{q}^Z_n$ be the largest symmetric equilibrium obedience rate induced by zero tolerance, where zero tolerance means

$p(r) = \mathbf{1}\{r \geq 1\}$. Then zero tolerance is obedience-maximizing if and only if $\widehat{q}_n^Z \geq 1 - \frac{1}{n}$. Unlike in the private-values benchmark, $\widehat{q}_n^Z$ depends on the payoff environment through G_n, so the finite zero-tolerance region is no longer summarized by the private-values number $\zeta(F)$. Assumption 4.12 keeps equilibrium obedience bounded away from zero and one, so the binomial tail bound and the central limit argument deliver the frequent-punishment conclusions.

The first-order stochastic dominance comparative static in Proposition 4.4(a) also continues to hold, since F enters the fixed-point map only through composition.

Moreover, the following monotonicity condition on the levels of obedience payoffs is useful for interpreting the equilibrium selection. Say that obedience has *no negative externalities* if, for every person i, every type θ, every own action $b \in \{0, 1\}$, every $j \neq i$, and every profile $a_{-i,-j} \in \{0, 1\}^{n-2}$,

$$w_i(b, a_{-i,-j}, 0, \theta) \geq w_i(b, a_{-i,-j}, 1, \theta).$$

Remark 4.15 (Obedience equilibria are Pareto ranked). Fix a weakly increasing p. If obedience has no negative externalities, then symmetric equilibria are weakly Pareto-ordered by their obedience rate.

In particular, the largest symmetric equilibrium yields everyone the maximal expected equilibrium payoff. The reason is that under an increasing punishment rule, higher obedience by others weakly lowers expected punishment under either own action. If obedience also has no negative externalities, the non-punishment payoff is weakly higher as well. Hence the value of best-responding is weakly increasing in the symmetric obedience rate.

4.3.4 Anonymous monitoring technologies

The baseline assumes exact-count monitoring: after actions are chosen, the ruler observes $|a|$ and conditions punishment directly on that count. I now relax that assumption: the ruler still does not observe identities, but instead observes a public signal drawn from an arbitrary

anonymous monitoring technology. The threshold logic then survives, now in signal space, provided the signal structure satisfies a simple monotonicity condition.

Let $S = \{s_1 < \cdots < s_K\}$ be a finite ordered signal space, and let $\alpha : \{0,1\}^N \to \Delta(S)$ be the conditional law of the public signal given the action profile. For a permutation $\sigma : N \to N$, write $a_i^\sigma := a_{\sigma(i)}$.

Definition 4.16. A monitoring technology α is *anonymous* if $\alpha(\cdot \mid a^\sigma) = \alpha(\cdot \mid a)$ for every action profile $a \in \{0,1\}^N$ and every permutation σ of N.

Exact count is the finest anonymous signal: the next lemma makes this precise.

Lemma 4.17 (Anonymous monitoring is a garbling of count)**.** *A monitoring technology α is anonymous if and only if there exist probability vectors $\pi_0, \ldots, \pi_n \in \Delta(S)$ such that*

$$\alpha(\cdot \mid a) = \pi_{|a|} \qquad \textit{for every } a \in \{0,1\}^N.$$

By Lemma 4.17, any anonymous monitoring technology can be represented by the family $\{\pi_r\}_{r=0}^n$, where $\pi_r(\cdot)$ is the distribution of the signal conditional on exactly r people disobeying. To recover a threshold characterization, I impose an ordering condition on these signal laws.

Definition 4.18. An anonymous monitoring technology is *monotone* if the family $\{\pi_r\}_{r=0}^n$ is monotone in the likelihood-ratio order: for every $r' > r$ and every $j' > j$,

$$\pi_{r'}(s_{j'})\pi_r(s_j) \geq \pi_{r'}(s_j)\pi_r(s_{j'}).$$

Higher signals are then evidence of higher disobedience. Exact-count monitoring satisfies this condition, and so do noisy count signals generated by many standard monotone likelihood-ratio error structures.

Fix a conjectured obedience rate $q \in (0,1)$. Let $B_q \sim \mathrm{Bin}(n, 1-q)$ be total disobedience, let $\psi_q(j) := \Pr(s = s_j)$ be the induced probability of signal s_j, and write $p_j := p(s_j)$ for a

signal rule $p : S \to [0,1]$. Using the pivotal-count weights $\Lambda_q(r)$ from (4.5), expected on-path punishment and deterrence are

$$\begin{aligned} P_p(q) &:= \sum_{j=1}^{K} \psi_q(j) p_j, \\ D_p(q) &:= \sum_{j=1}^{K} d_q(j) p_j, \\ d_q(j) &:= \sum_{r=0}^{n-1} \Lambda_q(r) \Big(\pi_{r+1}(s_j) - \pi_r(s_j) \Big). \end{aligned}$$

The key object is the deterrence-per-punishment ratio $d_q(j)/\psi_q(j)$. Monotone anonymous monitoring ranks signals by this ratio.

Lemma 4.19 (Signal ranking)**.** *Fix $q \in (0,1)$. Under monotone anonymous monitoring, $d_q(j)/\psi_q(j)$ is weakly increasing in j on $\{j : \psi_q(j) > 0\}$.*

Thus higher signals are cheaper sources of deterrence. The same exchange argument used for counts now implies threshold optimality in signal space.

Remark 4.20 (Fixed-q upper-threshold rules). Fix $q \in (0,1)$.

1. For any $\kappa \geq 0$, an optimizer of

$$\max_{0 \leq p_j \leq 1} \sum_{j=1}^{K} \Big(d_q(j) - \kappa \psi_q(j) \Big) p_j$$

   is an upper-threshold rule.

2. For any feasible target deterrence level $\overline{d}$, an optimizer of

$$\min_{0 \leq p_j \leq 1} \sum_{j=1}^{K} \psi_q(j) p_j \qquad \text{subject to} \qquad \sum_{j=1}^{K} d_q(j) p_j \geq \overline{d}$$

   is also an upper-threshold rule, with possible mixing at one boundary signal.

So the count model survives in form. Once q is fixed, the ruler's problem again reduces to choosing a cutoff, now in signal space rather than count space.

For $t \in \{1, \dots, K+1\}$, let $p_t(s_j) := \mathbf{1}\{j \geq t\}$, with $p_{K+1} \equiv 0$, and define $D_t(q) := D_{p_t}(q)$ and $D_S^*(q) := \max_{1 \leq t \leq K+1} D_t(q)$.

Proposition 4.21 (Anonymous monitoring)**.** *Let $T(q) := F(t_n(q, D_S^*(q)))$. Under Assumption 4.12 and monotone anonymous monitoring, every symmetric equilibrium obedience rate q implementable by some punishment rule satisfies $q \leq T(q)$. Consequently, the maximal implementable symmetric obedience rate is the largest fixed point of T, and it is attained by a threshold rule in signal space.*

The same logic also yields a natural posterior interpretation. Signals are valuable when they predict high realized disobedience.

Remark 4.22 (Posterior interpretation). For any signal s_j with $\psi_q(j) > 0$,

$$\frac{d_q(j)}{\psi_q(j)} = \frac{\mathbb{E}[B_q \mid s = s_j] - n(1-q)}{nq(1-q)}.$$

Thus a signal is more valuable for deterrence precisely when it predicts higher posterior disobedience.

The costly-punishment reduction also carries over with signals replacing counts. Suppose, as in Section 4.3.1, that the ruler's gross objective is obedience and that punishment has linear cost κ. Let $\tau_n(q)$ denote the least deterrence level d such that $q = F(t_n(q, d))$; in the private-values benchmark, $\tau_n(q) = F^{-1}(q)$. Define $C(q)$ as the least expected on-path punishment needed to deliver $\tau_n(q)$ units of deterrence:

$$C(q) := \min_{0 \leq p_j \leq 1} \sum_{j=1}^{K} \psi_q(j) p_j \quad \text{s.t.} \quad \sum_{j=1}^{K} d_q(j) p_j \geq \tau_n(q).$$

By Remark 4.20, whenever the constraint is feasible, an optimizer is upper-threshold in signal space, with possible mixing at one boundary signal. The same target-cost argument gives the reduced problem

$$\sup_{q : C(q) < +\infty} \{q - \kappa C(q)\}.$$

Under monotone anonymous monitoring, the comparative statics on the ruler's cost of punishment κ from Proposition 4.10 survive, with the boundary now in signal space. Moreover, if two anonymous monotone monitoring technologies are Blackwell ordered, maximal symmetric equilibrium obedience is weakly higher under the more informative technology.

Remark 4.23 (Comparative statics under anonymous monitoring). Let ℓ_κ denote the lowest punished signal index under an optimal upper-threshold rule. As κ increases, q_κ and P_κ weakly decrease, while ℓ_κ weakly increases. Moreover, if two anonymous monotone monitoring technologies are Blackwell ordered, then maximal symmetric equilibrium obedience is weakly higher under the more informative technology.

The welfare implications are less straightforward. A coarser anonymous signal weakly hurts the ruler: more informative monitoring can always be ignored. For people, the effect is ambiguous: coarser monitoring lowers the obedience burden, but it can also make punishment less precise and more frequent, because the ruler is less able to separate obedience from disobedience.

4.4 Conclusion

When a ruler observes only how many of his subjects disobey, he must choose where to place the punishment boundary against the distribution of disobedience that boundary itself induces. Optimal anonymous deterrence is *self-modal*: the ruler's tolerance coincides with a mode of the induced binomial distribution of others' disobedience. The ruler punishes the states in which an individual action is most likely to tip the group from safety into punishment. Groups that differ substantially in their propensity to obey can face similar equilibrium-path punishment rates, because the ruler adjusts the threshold to the distribution he faces. Repression frequency may thus be a poor indicator of the ruler's control over a population.

Appendix: Proofs

A.1 Proofs for Chapter 2

A.1.1 Proofs for Section 2.2

Lemma A.1. *For every finite x,*

$$\mathbb{E}\Big[(x-\theta_i)^+\Big] - \mathbb{E}\Big[(1-\theta_i)^+\Big] = J(x).$$

Proof. For every realization of θ_i, the difference on the left equals $\int_1^x \mathbf{1}\{\theta_i \le t\}\,dt$. Taking expectations and applying Fubini's theorem gives $\int_1^x F(t)\,dt = J(x)$. □

Derivation of Ann's continuation payoff and policing condition. Let $d_P := 1 - F(1+P\gamma)$. At type θ_A, Ann's optimized obedience-stage payoff is $\theta_A - 1 - \xi P d_P + \Big(x_A(r,P) - \theta_A\Big)^+$. Its expectation equals $\mathbb{E}[\theta_A - 1] + \mathbb{E}[(1-\theta_A)^+] + J\Big(x_A(r,P)\Big) - \xi P d_P$. The first two terms are independent of (r,P), while $P d_P = P\Big(1 - F(1+\gamma)\Big)$. This gives (2.2).

When $r = 0$, Ann's cutoff is 1 under either policing choice, so policing only adds its strictly positive expected cost. When $r = 1$, substituting $x_A(1,0) = F(1)$ and $x_A(1,1) = F(1+\gamma)$ gives (2.3). Its numerator is strictly positive because F and J are strictly increasing. As $\gamma \to 0$, the numerator converges to zero while the denominator converges to $1 - F(1) > 0$. □

Proof of Proposition 2.1. If (2.3) fails, the comparison in Section 2.2.4 shows that collective responsibility is strictly worse than individual responsibility. If it holds, $V_1 \ge V_0$ is equivalent to $\beta \ge \beta^*(\gamma)$. The inequalities $q_A^1 < F(1) < q_B^1$ imply that $\beta^*(\gamma) \in (0,1)$ and give the stated strictness conditions. □

Proof of Proposition 2.2. Continuity of $\mathscr{G}(\cdot;\gamma)$ makes $\mathcal{X}(\xi,\gamma)$ a closed subset of $[0,1]$. Moreover, $\mathscr{G}(0;\gamma) = 0 < \xi\Big(1 - F(1+\gamma)\Big)$. Thus, if $\mathcal{X}(\xi,\gamma)$ is nonempty, compactness gives a minimum $\underline{r} > 0$. Any $r > 0$ outside $\mathcal{X}(\xi,\gamma)$ fails to induce policing. Bob's cutoff is then 1, while Ann's cutoff is $1 - r(1 - F(1)) < 1$, so r is strictly dominated by individual responsibility. Within $\mathcal{X}(\xi,\gamma)$, Bob's cutoff remains $1+\gamma$, while Ann's cutoff $1 - r(1 - F(1+\gamma))$ strictly decreases in r. Because F is strictly increasing and $1 - \beta > 0$, the ruler's payoff also strictly decreases in r. He therefore prefers $\underline{r}$ to every other element of $\mathcal{X}(\xi,\gamma)$, so every optimum belongs to $\{0, \underline{r}\}$. □

A.1.2 Proofs for Section 2.3

A.1.2.1 Proofs for Section 2.3.2

Proof of Remark 2.3. Fix R and P, and let q_j denote the probability that person j obeys. By independence, given the strategies of the other people, person i's expected payoff gain from obeying is $\kappa \mathbf{1}\{iRi\} \prod_{j \ne i:\, iRj} q_j + \gamma d_i(P) - \theta_i$. This expression is strictly decreasing in θ_i, so i obeys if and only if her type is weakly below the cutoff given by its first two terms. The cutoff lies in $[0, \overline{\theta}]$. In a cutoff equilibrium, $q_j = F(x_j)$, so the cutoff vector satisfies (2.10). Conversely, if $x \in [0,\overline{\theta}]^n$ satisfies (2.10), then each person's gain from obeying when

the others use the cutoffs in x is $x_i - \theta_i$. The cutoff profile is therefore an equilibrium under the tie-breaking convention. □

Proof of Theorem 2.4. By Remark 2.3, equilibrium cutoff profiles are the fixed points of the map g defined in (2.10). This map sends $[0, \overline{\theta}]^n$ into itself and is weakly increasing in every coordinate. Tarski's fixed-point theorem therefore implies that its fixed points form a nonempty complete lattice and hence have smallest and largest elements.

The map g is pointwise weakly increasing in κ, γ, and P. Holding $R \setminus I$ fixed, enlarging $R \cap I$ turns on a nonnegative vertical-deterrence term in each affected row and therefore weakly raises g. Holding $R \cap I$ fixed, enlarging $R \setminus I$ adds factors in $[0, 1]$ to the vertical-deterrence term and therefore weakly lowers g. For any two environments ordered as in the theorem, let M be the larger of their feasible cutoff bounds. Both cutoff maps act on $[0, M]^n$ and are pointwise ordered in the stated direction. Monotone fixed-point comparison therefore gives all the stated comparative statics for both extremal equilibria. □

A.1.2.2 Proofs for Section 2.3.3

Lemma A.2 (Graph dependence of obedience cutoffs)**.** *Fix R and one of the two extremal obedience selections. For $r \neq j$ with $(r, j) \notin P$, define*

$$P^+ := P \cup \{(r, j)\}, \qquad x := x^*(R, P), \qquad x^+ := x^*(R, P^+). \tag{A.1}$$

Then, for every $h \in N$,

$$x_h^+ > x_h \quad \Longleftrightarrow \quad h = j \text{ or } h \overline{R}^+ j.$$

Consequently, for every $i \in N$,

$$\overline{\Phi}_i(R, P^+) < \overline{\Phi}_i(R, P) \quad \Longleftrightarrow \quad i \xrightarrow{R} j.$$

Proof. The added link raises only $d_j(P)$, so the cutoff map under P^+ weakly exceeds that under P. Theorem 2.4 therefore gives $x^+ \geq x$, while (2.10) gives $x_j^+ \geq x_j + \gamma > x_j$.

Suppose $h \overline{R}^+ j$ and take an $\overline{R}$-path from h to j. Starting from the strict increase in x_j, proceed backward along the path. If $h_t \overline{R} h_{t+1}$ and $x_{h_{t+1}}^+ > x_{h_{t+1}}$, then h_t is self-responsible and responsible for h_{t+1}. Equation (2.10) therefore gives $x_{h_t}^+ > x_{h_t}$ because F is strictly increasing and every other factor in the relevant product is at least $F(0) > 0$.

For the converse, let C contain the people other than j from whom j is not reachable through $\overline{R}$. If $h \in C$ and $h \overline{R} m$, then $m \in C$. Hence, for every $h \in C$, the coordinate g_h depends only on coordinates indexed by C, and the added policing link does not change d_h. Because g is continuous, its smallest and largest fixed points are obtained by monotone iteration from 0 and $\overline{\theta}\mathbf{1}$, respectively. The two iterations agree on C at every step, so $x_h^+ = x_h$ for every $h \in C$. This proves the cutoff claim.

Finally, $x^+ \geq x$ implies that i's punishment probability weakly falls. The fall is strict exactly when i is responsible for some ℓ whose cutoff rises strictly. By the cutoff claim and Definition 2.5, this is equivalent to $i \xrightarrow{R} j$. □

Lemma A.3 (Value effect of a policing link)**.** *Use the notation in* (A.1) *with* $r = i$*. Then*

$$V_i(R, P^+; \kappa, \gamma) > V_i(R, P; \kappa, \gamma) \quad \iff \quad i \overset{R}{\to} j.$$

If $i \overset{R}{\not\to} j$*, the two continuation values are equal.*

Proof. Integration by parts and the cutoff equation give the following representation:

$$\begin{aligned} H(z) &:= \int_{-\infty}^{z} F(t)\,dt, \\ V_i(R, P; \kappa, \gamma) &= H\Big(x_i^*(R, P)\Big) + B_i(P), \\ B_i(P) &:= \begin{cases} -\kappa - \gamma d_i(P), & \text{if } iRi, \\ -\kappa\overline{\Phi}_i(R, P) - \gamma d_i(P), & \text{if } (i,i) \notin R. \end{cases} \end{aligned} \tag{A.2}$$

Adding (i, j) does not change $d_i(P)$. If iRi and $i \overset{R}{\to} j$, then $i\overline{R}^+ j$, so Lemma A.2 gives $x_i^*(R, P^+) > x_i^*(R, P)$. Since H is strictly increasing, the continuation value rises strictly.

If $(i, i) \notin R$, then $x_i^*(R, P) = \gamma d_i(P)$ is unchanged. The continuation value therefore rises strictly exactly when $\overline{\Phi}_i$ falls strictly, which by Lemma A.2 occurs exactly when $i \overset{R}{\to} j$. When i is not connected to j, both the relevant cutoff and punishment probability are unchanged, so the continuation values are equal. □

Lemma A.4 (Uniform policing bounds)**.** *For fixed* R*,* P_{-i}*, and any two outgoing choices* P_i *and* P_i'*, the following bounds hold:*

$$\begin{aligned} \eta(\kappa, \gamma) &:= 1 - F\Big(\kappa + (n-1)\gamma\Big) > 0, \\ \Big|V_i\Big(R, (P_i, P_{-i}); \kappa, \gamma\Big) - V_i\Big(R, (P_i', P_{-i}); \kappa, \gamma\Big)\Big| &\le \kappa, \\ \xi\eta(\kappa, \gamma) \le \xi\Big(1 - F(x_j^*(R, P))\Big) &\le \xi. \end{aligned} \tag{A.3}$$

The final bounds apply to the expected cost of any policing link directed at j*.*

Proof. With outgoing policing costs suppressed, changing i's outgoing links affects her payoff only through ruler punishment. For every type and action, the resulting payoff difference is at most κ in absolute value. Taking maxima over actions and then expectations gives the second line of (A.3). The final line follows from $x_j^*(R, P) \in [0, \overline{\theta}]$. □

Proof of Theorem 2.6. Fix $\kappa, \gamma > 0$ and the selected extremal obedience continuation. Define

$$\widehat{\xi}(\kappa, \gamma) := \frac{\kappa}{\eta(\kappa, \gamma)}. \tag{A.4}$$

If $\xi > \widehat{\xi}(\kappa, \gamma)$, Lemma A.4 implies that every nonempty outgoing policing choice is strictly dominated by choosing no outgoing links. Hence no policing link can arise on path.

We next eliminate unconnected links. For fixed R, i, P_i, and P_{-i}, define

$$\begin{aligned} C_i^R &:= \{j \in N \setminus \{i\} : i \xrightarrow{R} j\}, \quad & P_i^c &:= P_i \cap C_i^R, \\ P &:= (P_i, P_{-i}), & P^c &:= (P_i^c, P_{-i}). \end{aligned} \tag{A.5}$$

By Lemma A.3, deleting the links in $P_i \setminus P_i^c$ leaves V_i unchanged. It also leaves the cutoff of every retained target unchanged. Indeed, if deleting a link directed at $j \notin C_i^R$ changed the cutoff of some $r \in C_i^R$, Lemma A.2 would imply either $r = j$ or $r \overline{R}^+ j$. The first is impossible, while the second would imply $i \xrightarrow{R} j$. Thus all retained-link costs are unchanged, while every deleted link has strictly positive cost. Hence P_i is strictly dominated by P_i^c whenever the two differ. No unconnected link can arise on path.

For every missing connected link define its gross gain and the minimum such gain by

$$\begin{aligned} G_i^j(R, P; \kappa, \gamma) &:= V_i(R, P \cup \{(i,j)\}; \kappa, \gamma) - V_i(R, P; \kappa, \gamma), \\ \Delta(\kappa, \gamma) &:= \min_{\substack{R, P,\ i \neq j: \\ (i,j) \notin P,\ i \xrightarrow{R} j}} G_i^j(R, P; \kappa, \gamma) > 0. \end{aligned} \tag{A.6}$$

Positivity follows from Lemma A.3 and finiteness of the sets of responsibility and policing relations. Define

$$\underline{\xi}(\kappa, \gamma) := \frac{\Delta(\kappa, \gamma)}{2}, \qquad \overline{\xi}(\kappa, \gamma) := \max\Big\{\widehat{\xi}(\kappa, \gamma), \underline{\xi}(\kappa, \gamma)\Big\}. \tag{A.7}$$

If $\xi \leq \underline{\xi}(\kappa, \gamma)$, adding any missing connected link gives a gross gain of at least $\Delta(\kappa, \gamma)$. The new link costs at most ξ, while the expected costs of existing links weakly fall because the added link weakly raises every selected cutoff. The net gain is therefore at least $\Delta(\kappa, \gamma) - \xi > 0$. Every connection between distinct people must consequently be present on path. Together with the preceding arguments, this proves both parts of the theorem. □

Lemma A.5 (Increasing effects of policing)**.** *Suppose that F is convex on the cutoff ranges being compared. Fix R and the selected extremal obedience equilibrium, and define*

$$x(P) := x^*(R, P), \qquad q_h(P) := F(x_h(P)). \tag{A.8}$$

For any $P' \supseteq P$ and any policing link $e \notin P'$,

$$\begin{aligned} x(P' \cup \{e\}) - x(P') &\geq x(P \cup \{e\}) - x(P), \\ q_h(P' \cup \{e\}) - q_h(P') &\geq q_h(P \cup \{e\}) - q_h(P) \qquad \textit{for every } h. \end{aligned}$$

These increments are also weakly increasing in κ and in γ.

Proof. Suppose $x' \geq x$ and $z' \geq z \geq 0$. Convexity and monotonicity of F imply that

$$F(x'_\ell + z'_\ell) - F(x'_\ell) \geq F(x_\ell + z_\ell) - F(x_\ell) \geq 0$$

for every ℓ. Writing each factor as its baseline value plus its increment and expanding products therefore gives, for every $A \subseteq N$,

$$\prod_{\ell \in A} F(x'_\ell + z'_\ell) - \prod_{\ell \in A} F(x'_\ell) \geq \prod_{\ell \in A} F(x_\ell + z_\ell) - \prod_{\ell \in A} F(x_\ell). \tag{A.9}$$

Let g^Q denote the cutoff map in (2.10) under policing relation Q, and write $e = (r, j)$. For the smallest fixed points, define four iterations from zero:

$$x_0^{00} = x_0^{10} = x_0^{01} = x_0^{11} = 0,$$

and, for every $t \geq 0$,

$$x_{t+1}^{00} = g^P(x_t^{00}), \qquad x_{t+1}^{10} = g^{P \cup \{e\}}(x_t^{10}), \qquad x_{t+1}^{01} = g^{P'}(x_t^{01}), \qquad x_{t+1}^{11} = g^{P' \cup \{e\}}(x_t^{11}).$$

We show by induction that

$$x_t^{01} \geq x_t^{00}, \qquad x_t^{10} \geq x_t^{00}, \qquad x_t^{11} \geq x_t^{01}, \qquad x_t^{11} - x_t^{01} \geq x_t^{10} - x_t^{00}. \tag{A.10}$$

The claims hold at $t = 0$. Suppose they hold at t. The first three inequalities at $t+1$ follow from $P' \supseteq P$, the pointwise ordering of the corresponding cutoff maps, and their monotonicity in x. For the final inequality, let $A_h := \{\ell \neq h : hR\ell\}$ and $s_h := \mathbf{1}\{hRh\}$. Because adding e raises only d_j by one,

$$x_{t+1,h}^{11} - x_{t+1,h}^{01} = \kappa s_h \left[\prod_{\ell \in A_h} F(x_{t,\ell}^{11}) - \prod_{\ell \in A_h} F(x_{t,\ell}^{01}) \right] + \gamma \mathbf{1}\{h = j\},$$

whereas

$$x_{t+1,h}^{10} - x_{t+1,h}^{00} = \kappa s_h \left[\prod_{\ell \in A_h} F(x_{t,\ell}^{10}) - \prod_{\ell \in A_h} F(x_{t,\ell}^{00}) \right] + \gamma \mathbf{1}\{h = j\}.$$

Apply (A.9) with

$$x' = x_t^{01}, \qquad z' = x_t^{11} - x_t^{01}, \qquad x = x_t^{00}, \qquad z = x_t^{10} - x_t^{00}.$$

The induction hypothesis gives $x' \geq x$ and $z' \geq z \geq 0$, so the first difference is weakly larger than the second in every coordinate. This proves (A.10) at $t+1$. Taking limits yields

$$x^{\min}(R, P' \cup \{e\}) - x^{\min}(R, P') \geq x^{\min}(R, P \cup \{e\}) - x^{\min}(R, P).$$

For the largest fixed points, initialize all four iterations at the common upper bound $\overline{\theta}\mathbf{1}$ and use the same four recursions. The iterations decrease to the corresponding largest fixed points. The induction above applies without change, since the four initial vectors are again equal and the argument uses only the ordering of the maps and (A.9). Taking limits gives the same inequality for the largest equilibrium selection.

The comparison across values of κ is identical. Fix $\kappa' \geq \kappa$ and compare the four maps associated with (κ, P), $(\kappa, P \cup \{e\})$, (κ', P), and $(\kappa', P \cup \{e\})$. The baseline map is weakly

higher under κ'. In the increment comparison, the product difference under κ' is weakly larger by (A.9) and is multiplied by the weakly larger coefficient κ'. Iteration from zero gives the result for the smallest fixed points, while iteration from the common upper bound $\Big(\kappa' + (n-1)\gamma\Big)\mathbf{1}$ gives it for the largest fixed points.

Now fix $\gamma' \geq \gamma$ and compare the four maps associated with (γ, P), $(\gamma, P \cup \{e\})$, (γ', P), and $(\gamma', P \cup \{e\})$. The baseline map is weakly higher under γ'. The propagated product difference is weakly larger by (A.9), while the direct increment in coordinate j is γ' rather than γ. The same induction from zero and from the common upper bound $\Big(\kappa + (n-1)\gamma'\Big)\mathbf{1}$ proves that the cutoff increments are weakly increasing in γ.

Finally, apply convexity and monotonicity of F to the ordered baseline cutoffs and cutoff increments already established. For every h,

$$F\Big(x_h(P' \cup \{e\})\Big) - F\Big(x_h(P')\Big) \geq F\Big(x_h(P \cup \{e\})\Big) - F\Big(x_h(P)\Big),$$

and the same argument gives the parameter comparisons for the obedience-probability increments. □

Proof of Proposition 2.8. Fix R, κ, γ, ξ and the selected extremal obedience equilibrium. Using the notation in (A.2) and (A.8), define

$$\begin{aligned} R_i &:= \{h \in N : iRh\}, & O_i(P) &:= \{r \in N : iPr\}, \\ C_i(P) &:= -\xi \sum_{r \in O_i(P)} \Big(1 - q_r(P)\Big), & \widetilde{U}_i^R(P) &= H(x_i(P)) + B_i(P) + C_i(P). \end{aligned} \tag{A.11}$$

For an outgoing link $e = (i, j)$, define

$$\Delta_e Z(P) := Z(P \cup \{e\}) - Z(P). \tag{A.12}$$

Fix $P' \supseteq P$ with $e \notin P'$. We show that $\Delta_e \widetilde{U}_i^R(P') \geq \Delta_e \widetilde{U}_i^R(P)$. If iRi, then $\Delta_e B_i(P') = \Delta_e B_i(P) = 0$. If $(i,i) \notin R$, the only component of B_i affected by e is $\kappa \prod_{h \in R_i} q_h(P)$. Lemma A.5 and (A.9) therefore imply that $\Delta_e B_i(P') \geq \Delta_e B_i(P)$.

Because $H' = F$, the function H is increasing and convex. Lemma A.5 therefore gives $\Delta_e H(x_i(P')) \geq \Delta_e H(x_i(P))$.

Finally,

$$\Delta_e C_i(P) = -\xi + \xi q_j(P \cup \{e\}) + \xi \sum_{r \in O_i(P)} \Big(q_r(P \cup \{e\}) - q_r(P)\Big).$$

Since $O_i(P') \supseteq O_i(P)$, Lemma A.5 implies that $\Delta_e C_i(P') \geq \Delta_e C_i(P)$. Adding the three comparisons gives the required increasing-differences inequality. Thus each person's payoff is supermodular in her own policing links and has increasing differences between her own links and the full policing relation. The induced policing game is supermodular and, because each action set is a finite lattice, has least and greatest pure equilibria.

The same comparisons, together with the parameter part of Lemma A.5, show that the marginal payoff from every outgoing link is weakly increasing in κ and in γ. Topkis's mono-

tone comparative-statics theorem therefore implies that both extremal policing equilibria are weakly increasing in these parameters. □

A.1.2.3 Proofs for Section 2.3.4

Lemma A.6 (Cutoff dominance)**.** *If two induced outcomes have cutoff profiles $x' \geq x$, then the ruler's expected value is weakly higher under x'; strictly so if $x'_i > x_i$ for some i.*

Proof. Couple the outcomes using common independent uniform random variables, with person i obeying under cutoff profile x if and only if her draw is at most $F(x_i)$, and write $a(x)$ for the resulting action profile. If $x' \geq x$, then $a(x') \leq a(x)$ coordinatewise almost surely. If some cutoff is strictly higher, strict monotonicity of F implies that the inequality is strict in that coordinate with positive probability. The result follows because v is strictly decreasing in each coordinate. □

Proof of Proposition 2.9. Suppose first that $R \neq I$ induces no policing. Individual responsibility also induces no policing because it has no cross-person connections. By (2.10), $x_i(I) = \kappa$ and $x_i(R) \leq \kappa$ for every i. The inequality is strict for at least one person: if some i is not self-responsible, then $x_i(R) = 0$; otherwise, some self-responsible row contains an off-diagonal link, and its cutoff is strictly below κ because the corresponding factor $F(x_j(R))$ is strictly below one. Lemma A.6 therefore gives $V(I) > V(R)$.

If $\xi > \overline{\xi}(\kappa, \gamma)$, Theorem 2.6 implies that every responsibility relation induces no policing, so the preceding argument makes I uniquely optimal. Finally, fix $\xi, \gamma > 0$. By (A.3), the gross continuation benefit from any nonempty outgoing policing choice is at most κ, whereas its expected cost is at least $\xi[1 - F(\kappa + (n-1)\gamma)]$. As $\kappa \downarrow 0$, the latter converges to the strictly positive number $\xi[1 - F((n-1)\gamma)]$. Hence every nonempty outgoing choice is strictly dominated when κ is sufficiently small. Policing is then empty under every responsibility relation, and I is again uniquely optimal. □

Proof of Proposition 2.10. Because policing is cheap, Theorem 2.6 implies that iPj if and only if $i \xrightarrow{R} j$ for every $i \neq j$.

Suppose first that an optimal R is not reflexive, and let $R' := R \cup \{(i,i)\}$ for some i who is not self-responsible. Adding (i,i) weakly expands connection and therefore policing, while turning on a strictly positive vertical-deterrence term for i. Theorem 2.4 gives $x(R') \geq x(R)$, with a strict inequality for i, contradicting Lemma A.6. Hence every optimal relation is reflexive.

Now suppose that $(i,j) \in R \setminus I$ is unnecessary for connection, and let $R' := R \setminus \{(i,j)\}$. Connection, and hence policing, is unchanged. Because R is reflexive, removing (i,j) removes a factor strictly below one from person i's vertical-deterrence term. Theorem 2.4 again gives $x(R') \geq x(R)$, with a strict inequality for i, contradicting Lemma A.6. Every optimal relation is therefore connection-minimal, proving part (a).

For part (b), fix $\gamma \in \Gamma_{\kappa,\xi}$ and a symmetric relation R. Its connection relation is symmetric, and every symmetry of R maps an equilibrium cutoff profile into another. The selected extremal profile is therefore invariant under every symmetry, so all people have a common cutoff x_R. Let x_C denote the selected common cutoff under a cycle-responsibility rule. By

symmetry, the following quantities are independent of the person i used to define them:

$$\begin{aligned} s_R &:= \mathbf{1}\{iRi\}, & r_R &:= \big|\{j \neq i : iRj\}\big|, & c_R &:= \big|\{j \neq i : j \xrightarrow{R} i\}\big|, \\ g_R(z) &:= \kappa s_R F(z)^{r_R} + \gamma c_R, & g_C(z) &:= \kappa F(z) + \gamma(n-1), & & \\ x_R &= g_R(x_R), & x_C &= g_C(x_C). & & \end{aligned} \tag{A.13}$$

Individual responsibility gives every person cutoff κ, so if $x_R < \kappa$, it strictly dominates R.

Suppose instead that $x_R \geq \kappa$ and $R \notin \{I\} \cup \mathcal{C}$. By (A.13),

$$g_C(x_R) - x_R = \kappa\Big[F(x_R) - s_R F(x_R)^{r_R}\Big] + \gamma(n - 1 - c_R). \tag{A.14}$$

If $s_R = 0$, the first term is strictly positive. If $s_R = 1$, then $R \neq I$ implies $r_R \geq 1$, and both terms in (A.14) are nonnegative. Equality requires $r_R = 1$ and $c_R = n - 1$. In that case the off-diagonal responsibility graph has one outgoing link per person and is strongly connected, so it is a single directed cycle and $R \in \mathcal{C}$, contrary to hypothesis. Thus $g_C(x_R) > x_R$.

Moreover, $g_C(z) \geq g_R(z)$ for every z. Monotone fixed-point comparison gives $x_C \geq x_R$ for the selected extremal solutions. Equality would imply $g_C(x_R) = x_R$, contradicting the preceding strict inequality. Hence $x_C > x_R$, and Lemma A.6 implies that a cycle-responsibility rule strictly dominates R. This proves part (b).

It remains to compare individual and cycle responsibility as γ varies. For every $\gamma \geq 0$, let $\widehat{x}_C(\gamma)$ be the selected extremal solution of

$$\widehat{x}_C(\gamma) = \kappa F\Big(\widehat{x}_C(\gamma)\Big) + \gamma(n - 1). \tag{A.15}$$

The map in (A.15) shifts strictly upward with γ. Monotone fixed-point comparison on a common bounding interval therefore implies that $\widehat{x}_C(\gamma)$ is strictly increasing.

At $\gamma = 0$, every solution of (A.15) is strictly below κ. Moreover, every solution satisfies $x \leq \kappa + \gamma(n - 1)$; by continuity, $\kappa F(\kappa + \gamma(n - 1)) + \gamma(n - 1) < \kappa$ for all sufficiently small positive γ, so every solution remains below κ near zero. Conversely, every solution exceeds κ whenever $\gamma(n - 1) > \kappa$. Define

$$\gamma^* := \sup\Big\{\gamma \geq 0 : \widehat{x}_C(\gamma) < \kappa\Big\}. \tag{A.16}$$

Then $0 < \gamma^* < \infty$. Strict monotonicity implies that $\widehat{x}_C(\gamma) < \kappa$ when $\gamma < \gamma^*$ and $\widehat{x}_C(\gamma) > \kappa$ when $\gamma > \gamma^*$, with no restriction at equality.

For $\gamma \in \Gamma_{\kappa,\xi}$, the actual cutoff under cycle responsibility equals $\widehat{x}_C(\gamma)$. Lemma A.6 therefore implies that individual responsibility is uniquely optimal among symmetric relations when $\gamma < \gamma^*$, while only cycle-responsibility relations are optimal when $\gamma > \gamma^*$. □

Proof of Remark 2.11. For fixed R, Proposition 2.8 implies that the selected extremal policing relation is weakly increasing in κ and γ. Theorem 2.4 then implies that the selected cutoff profile, and hence $V(R)$ by Lemma A.6, is weakly increasing in either parameter. Taking the maximum over responsibility relations preserves this monotonicity.

Under individual responsibility, there are no cross-person connections, policing is empty, and every cutoff equals κ, so $V(I)$ is independent of γ. If I is optimal at some γ', then for

every $\gamma < \gamma'$ and every R, monotonicity gives $V(R;\gamma) \le V(R;\gamma') \le V(I;\gamma') = V(I;\gamma)$. The set of values at which I is optimal is therefore a lower interval. □

Proof of Proposition 2.12. By (2.10), every cutoff is at most $\overline{\theta}$. Fix i and consider the star S_i, defined in the main text. Person i is self-responsible and has no off-diagonal responsibility links, while every $j \neq i$ is directly connected to i. Theorem 2.6 therefore implies that every $j \neq i$ polices i, giving $x_i = \kappa + (n-1)\gamma = \overline{\theta}$.

Suppose that distinct people i and j both attain $\overline{\theta}$. Each must have vertical-deterrence term κ and $n-1$ incoming policing links. Thus each is self-responsible and has no off-diagonal responsibility links, while jPi. But a person with no off-diagonal responsibility links cannot be connected to anyone else. This contradicts Theorem 2.6, which requires $j \overset{R}{\to} i$ whenever jPi. Hence no two distinct people can both attain $\overline{\theta}$. □

A.1.3 Proofs for Section 2.4

A.1.3.1 Proofs for Section 2.4.1

The monotonicity claim follows from (2.14). If $q_B(1) \ge q_B(0)$, then the difference of the two positive-part terms is weakly decreasing in θ_A, and strictly decreasing between $\kappa q_B(0)$ and $\kappa q_B(1)$ when the inequality is strict. The cost term is independent of θ_A, so the types who weakly prefer policing form a lower interval.

Proof of Proposition 2.13. Consider the stated assessment. Given $q_A(1) = 1$ and $q_A(0) = 0$, Bob's cutoffs are $\kappa + \gamma$ and 0, so his obedience probabilities are $F(\kappa + \gamma)$ and $F(0)$. By (2.15), $\xi < \xi^*$ implies $x_P^* > \kappa F(\kappa)$, while $\xi > 0$ implies $x_P^* < \kappa F(\kappa + \gamma)$. Hence every type who polices obeys after policing, establishing $q_A(1) = 1$.

After no policing, any candidate posterior satisfies $q_A(0) \le 1$, so Bob's cutoff is at most κ and his obedience probability is at most $F(\kappa)$. Ann's obedience cutoff is therefore at most $\kappa F(\kappa) < x_P^*$. Every type in the no-policing pool disobeys, establishing $q_A(0) = 0$ and hence Bob's cutoff $x_B(0) = 0$. This also shows that the no-policing continuation is unique.

Substituting these continuation probabilities into (2.14), the policing gain crosses zero at x_P^* by (2.15). Because x_P^* lies strictly between $\kappa F(0)$ and $\kappa F(\kappa + \gamma)$, the piecewise-linear expression is strictly positive below x_P^* and strictly negative above it. Ann therefore polices exactly when $\theta_A \le x_P^*$. Finally, strict increase of F on $\mathbb{R}$ implies $F(x_P^*) \in (0,1)$, so both policing decisions occur with positive probability and the stated posteriors are Bayes-consistent. The assessment is therefore a pure-strategy PBE. □

A.1.3.2 Proofs for Section 2.4.2

Proof of Remark 2.14. In (2.16), neither person gains by deviating from CC, while a unilateral deviation from DD yields $-\eta < 0$. Thus both outcomes are continuation equilibria. Because admissible maps satisfy $\pi_{ij}(0) = CC$, each map is determined by whether $\pi_{ij}(1) = CC$ or $\pi_{ij}(1) = DD$, giving the stated bijection with irreflexive policing relations. Relative to CC, switching to DD after j disobeys imposes losses ξ on i and γ on j. Summing across ordered pairs gives exactly the policing-cost and peer-sanction terms in (2.9), so the obedience and policing incentives coincide with those in the baseline model. □

Proof of Proposition 2.15. Fix R and a. If $i \notin \mathscr{S}(R,a)$, her own disobedience makes ruler punishment unavoidable, so no outgoing link saves her punishment and she sanctions nobody. Suppose instead that $i \in \mathscr{S}(R,a)$. She avoids ruler punishment if and only if she sanctions every member of $D_i(R,a)$. Any proper nonempty subset therefore adds sanctioning costs without preventing punishment and is strictly dominated by sanctioning nobody. Person i consequently chooses between paying κ and sanctioning all $m_i(R,a)$ relevant offenders at cost $\xi m_i(R,a)$. When $\xi m_i(R,a) = \kappa$, sanctioning everyone and sanctioning nobody are both optimal. Selecting sanctioning everyone at equality gives the equilibrium stated in the proposition. □

Derivation of equation (2.18). Fix the other people's strategies and a_{-i}. If i obeys, her selected cost from ruler punishment and outgoing sanctions is $\min\{\xi m_i(R,a_{-i}),\kappa\}$. If she disobeys, this cost is unchanged when she is not self-responsible and equals κ when she is self-responsible. Disobedience also exposes her to the incoming sanctions counted by $s_i(R,a_{-i})$. Her expected gain from obeying is therefore the right-hand side of (2.18) minus θ_i. It is strictly decreasing in θ_i, giving the stated cutoff equation.

Let T denote the right-hand side of (2.18). It maps $[0,\overline{\theta}]^n$ into itself. To see that it is increasing, couple the action profiles generated by two ordered cutoff vectors so that an increase in a cutoff can only change a person's action from disobedience to obedience. The number $m_i(R,a_{-i})$ then weakly falls, weakly raising the first term of T_i. For every $h \neq i$, greater obedience by the other people weakly expands $\mathscr{S}(R,a)$ and weakly lowers $m_h(R,a)$, so the indicator that h sanctions i weakly rises. The second term of T_i therefore also weakly rises. Thus T is increasing, and Tarski's fixed-point theorem gives smallest and largest cutoff equilibria. □

Proof of Remark 2.16. Fix i and use the star S_i. Person i is self-responsible, while every $h \neq i$ is responsible only for i. Whenever i disobeys, each such h faces one relevant offender and sanctions because $\xi \leq \kappa$. Person i therefore faces her own ruler punishment and all $n-1$ peer sanctions, giving $x_i = \overline{\theta}$.

Now suppose that distinct people i and j both attain $\overline{\theta}$. Maximality of x_i requires j to sanction i with probability one conditional on i's disobedience. Maximality of x_j requires the full direct incentive κ and hence jRj. Because $F(\overline{\theta}) < 1$, conditional on i's disobedience, j also disobeys with positive probability. On that event, $j \notin \mathscr{S}(R,a)$ and Proposition 2.15 implies that she imposes no outgoing sanctions. This contradicts the requirement that she always sanction i following i's disobedience. □

A.1.3.3 Proofs for Section 2.4.3

Proof of Proposition 2.17. Given the other people's strategies, person i's expected gain from obeying is the right-hand side of (2.20) minus θ_i. It is strictly decreasing in θ_i, so every equilibrium is in cutoff strategies and its cutoff vector satisfies (2.20); the converse follows from the same best-response calculation.

The right-hand side of (2.20) lies in $[0, M_L]$ and is weakly increasing in every cutoff. Tarski's fixed-point theorem therefore gives smallest and largest fixed points. The cutoff map is pointwise weakly increasing in κ, γ, λ, L, and P. Comparing any two environments

on a common bounding lattice gives the stated comparative statics for both extremal fixed points. □

Proof of Remark 2.18. If $L = \varnothing$, (2.20) reduces to (2.10), so the claims follow from Theorem 2.4. Conversely, suppose iLj. With $P = \varnothing$, compare the empty responsibility relation with $\{(j, i)\}$. All cutoffs are zero under the former, whereas the latter gives $x_i = \lambda$ and leaves every other cutoff at zero. Thus obedience need not decrease when an off-diagonal responsibility edge is added.

Next compare $\{(j, i)\}$ with $\{(j, i), (j, j)\}$. Under the first relation, $x_i = \lambda$. Under the second, every equilibrium satisfies $x_i = \lambda F(x_j) < \lambda$. Thus obedience need not increase when a self-responsibility edge is added. Either comparative static can therefore hold uniformly over responsibility relations only if $L = \varnothing$. □

Lemma A.7 (Altruistic graph dependence)**.** *Fix R, L, and one of the two extremal obedience selections. For $r \neq j$ and $(r, j) \notin P$, define*

$$P^+ := P \cup \{(r, j)\}, \qquad x := x^*(R, P), \qquad x^+ := x^*(R, P^+). \tag{A.17}$$

Then $x_h^+ > x_h$ if and only if $h = j$ or $h\widehat{R}^+ j$. Moreover, $C_i(x^+) < C_i(x)$ if and only if $i \overset{R}{\rightsquigarrow} j$.

Proof. For distinct h and m, the right-hand side of (2.20) is strictly increasing in x_m if and only if $h\widehat{R}m$. Adding (r, j) directly raises coordinate j by γ. The monotone-iteration proof of Lemma A.2 therefore applies with $\widehat{R}$ in place of $\overline{R}$, proving the first claim.

Because $x^+ \geq x$, every punishment probability internalized in C_i weakly falls. The fall is strict exactly when some person whose punishment i internalizes is responsible for a person whose cutoff rises strictly. By the first claim and (2.22), this is equivalent to $i \overset{R}{\rightsquigarrow} j$. □

For use below, let $V_i^L(R, P)$ denote person i's ex ante continuation payoff with outgoing policing costs set equal to zero. Her payoff in the induced policing game is

$$\widetilde{U}_i^{L,R}(P) := V_i^L(R, P) - \xi \sum_{j:\, iPj} \Big(1 - F(x_j^*(R, P))\Big). \tag{A.18}$$

Lemma A.8 (Altruistic value effect of a policing link)**.** *In the setting of Lemma A.7, with $r = i$, $V_i^L(R, P^+) > V_i^L(R, P)$ if and only if $i \overset{R}{\rightsquigarrow} j$. Otherwise the two continuation values are equal.*

Proof. Couple obedience under P and P^+ using common type realizations. Lemma A.7 implies that every person who obeys under P also obeys under P^+. For every type of i and each fixed action, this weakly lowers the punishment costs internalized by i, while her incoming peer sanctions are unchanged. Maximizing over her action therefore gives $V_i^L(R, P^+) \geq V_i^L(R, P)$.

If $i \overset{R}{\not\rightsquigarrow} j$, Lemma A.7 implies that none of the punishment probabilities internalized by i changes, so equality holds. Suppose instead that $i \overset{R}{\rightsquigarrow} j$. If some cutoff other than i's that enters a punishment probability internalized by i rises strictly, sufficiently low types of i obey in both environments and receive a strict reduction in expected punishment costs.

Otherwise, the definition of $\overset{R}{\rightsquigarrow}$ and Lemma A.7 imply that i's own cutoff rises strictly. Strict increase of F then gives positive probability to types between the old and new cutoffs. These types switch from disobedience to obedience and gain strictly. In either case the ex ante continuation value rises strictly. □

Proof of Proposition 2.19. Every cutoff is at most M_L, so each policing link has expected cost at least $\xi[1-F(M_L)] > 0$. For fixed κ, γ, λ, and L, there are finitely many responsibility and policing relations, and all continuation values are finite. Gross continuation gains from outgoing policing choices are therefore uniformly bounded. For sufficiently large ξ, every nonempty outgoing choice is strictly dominated by choosing no outgoing links. This gives a finite upper threshold above which policing is absent.

Next fix R and i. From any outgoing choice, delete every link (i,j) for which $i \overset{R}{\not\rightsquigarrow} j$. Lemma A.8 implies that these deletions leave V_i^L unchanged. They also leave the costs of retained links unchanged. Indeed, if deleting (i,j) changed the cutoff of a retained target r, Lemma A.7 would imply either $r = j$ or $r\widehat{R}^+ j$. The first is impossible, while the second, together with $i \overset{R}{\rightsquigarrow} r$, would imply $i \overset{R}{\rightsquigarrow} j$. Each deleted link has strictly positive expected cost, so any outgoing choice containing such a link is strictly dominated. Hence iPj on path only if $i \overset{R}{\rightsquigarrow} j$.

Finally, Lemma A.8 gives a strictly positive gross continuation gain from adding any missing link (i,j) with $i \overset{R}{\rightsquigarrow} j$. There are finitely many such configurations, so these gains have a strictly positive minimum. Adding the link costs at most ξ, while the expected costs of existing links weakly fall because all selected cutoffs weakly rise. For sufficiently small positive ξ, every such addition is therefore strictly profitable. Thus every on-path relation satisfies $P = \overset{R}{\rightsquigarrow} \setminus I$. Choosing the lower and upper thresholds, and enlarging the latter if necessary, gives $0 < \underline{\xi} \leq \overline{\xi}$ and both conclusions of the proposition. □

Proof of Remark 2.20. If $\xi > \overline{\xi}$, Proposition 2.19 implies that policing is absent. Starting from any responsibility relation, delete every edge outside $I \cup L^{-1}$. Consider one such edge (h,k). Then $h \neq k$ and $k \not L h$. Deleting the edge weakly raises h's self-pivotality term by removing a factor $F(x_k)$. It also cannot destroy an altruistic-pivotality term: the only such term that the edge itself could create is the term for person k generated by kLh, which is ruled out. In every existing altruistic-pivotality term involving h's punishment, deletion can only remove the factor $F(x_k)$.

The cutoff map therefore rises pointwise after each deletion. Monotone fixed-point comparison implies that both extremal cutoff profiles weakly rise, and the common-uniform coupling then implies that the ruler's value weakly rises. The ruler can consequently restrict attention to relations contained in $I \cup L^{-1}$. □

A.1.3.4 Proofs for Section 2.4.4

Proof of Proposition 2.21. Fix R and P. Given cutoff x_i, person i's good-signal probability is $1-\alpha-\delta(1-F(x_i))$. Obeying lowers her bad-signal probability by δ, so equilibrium cutoffs

satisfy

$$x_i = \kappa\delta\,\mathbf{1}\{iRi\}\prod_{\substack{j\neq i\\ iRj}}\Big(1-\alpha-\delta(1-F(x_j))\Big)+\gamma d_i(P). \tag{A.19}$$

Every cutoff lies in $[0,\overline{\theta}]$. For every type of i and each fixed action, changing her outgoing policing links affects only her ruler-punishment probability because her incoming peer sanctions are fixed. Each good-signal factor lies in an interval of width δ, so a telescoping product comparison bounds the change in this punishment probability by $n\delta$. Maximizing over i's action, the gross continuation benefit of any outgoing choice is therefore at most $\kappa n\delta$. Each policing link has expected cost at least $\xi[1-F(\overline{\theta})]>0$. Choose $\overline{\delta}\leq 1-\alpha$ sufficiently small that $\kappa n\overline{\delta}<\xi[1-F(\overline{\theta})]$. For every $\delta<\overline{\delta}$, every nonempty outgoing choice is strictly dominated and policing is absent under every responsibility relation.

Under individual responsibility, (A.19) gives $x_i(I)=\kappa\delta$ for every i. Under any $R\neq I$, every cutoff is weakly below $\kappa\delta$. The inequality is strict for at least one person: a person who is not self-responsible has cutoff zero, while a self-responsible person with an off-diagonal responsibility link has a cutoff strictly below $\kappa\delta$ because the corresponding good-signal factor is strictly below one. The common-uniform coupling therefore gives $V(I)>V(R)$. Hence individual responsibility is uniquely optimal whenever $0<\delta<\overline{\delta}$. □

A.2 Proofs for Chapter 3

A.2.1 Proofs for Section 3.2.2

Proof of Theorem 3.3. Fix a denunciation regime $\mathcal{D}$ and suppose that S is publicly observed before messages are chosen. Let $B\in\mathcal{D}$ be a block and write $C:=S\cap B$.

First consider equilibrium messages. A message naming someone outside B cannot affect whether B is collectively punished and cannot affect whether the sender is targeted inside B. A message naming someone in $B\setminus C$ cannot help satisfy the coverage condition $S\cap B\subseteq M$. A self-report can only add the sender to the targeted punished set and never helps the sender avoid collective punishment. Since every reported name costs $\xi>0$, every message used with positive probability satisfies $M_i\subseteq S\cap B\setminus\{i\}$.

It remains to compare singleton and nonsingleton regimes. Under the singleton regime, person i is punished if and only if she disobeys and is detected. Thus the singleton cutoff is $x_1=\mu$.

Now consider a nonsingleton block B and fix $i\in B$. With public detection, undetected disobedience has no reporting role: reports can cover only the public detected set. Conditional on i being undetected, her reporting continuation value is the same whether she obeyed or disobeyed. If i is detected, she is punished regardless of her message. Hence the marginal punishment benefit of obedience is at most μ.

It is strictly below μ in every nonsingleton block. Since $F(1)<1$, each other member of B disobeys with positive probability even under cutoff one. With positive probability, some other member of B disobeys and is detected. In that event, an undetected person in B either bears reporting costs to cover detected disobedience or faces collective punishment if detected disobedience is left uncovered. This continuation cost is borne whether i obeyed or

disobeyed without being detected. Therefore the cutoff in any nonsingleton block is strictly below the singleton cutoff μ.

Thus the singleton regime induces weakly more obedience for every person and strictly more obedience for every person who would otherwise belong to a nonsingleton block. Since $v(a)$ is strictly decreasing in each coordinate, the singleton regime uniquely maximizes $\mathbb{E}[v(a)]$ under symmetric equilibria. □

A.2.2 Proofs for Section 3.2.3

Proof of Lemma 3.4. Fix a partition $\mathcal{D}$, a realized disobedience set A, a block $B \in \mathcal{D}$, and a person $i \in B$. People observe the realized disobedience set A, but a message can affect person i's payoff only through punishment in her own block B.

A message naming someone outside B cannot affect whether B is collectively punished and cannot affect whether i is individually targeted in B. A message naming someone who obeyed cannot help cover detected disobedience, because obedient people are never detected. A self-report is also dominated. If i obeyed, self-reporting names an obedient person and is useless. If i disobeyed, then when i is detected she is punished whether or not she names herself; when she is not detected, naming herself can only put her in the targeted punished set. In all cases, deleting such a name weakly lowers punishment risk and strictly lowers reporting cost.

Hence every message used with positive probability satisfies $M_i \subseteq A \cap B \setminus \{i\}$. □

Proof of Lemma 3.5. Fix a block B, a realized disobedience set A, and a person $i \in B$. By Lemma 3.4, the only names that can appear in an undominated message are the names in $M_i^{\max} := A \cap B \setminus \{i\}$. Let $K := |M_i^{\max}|$. If $K \leq 1$, there is no interior report size. Suppose $K \geq 2$.

Under symmetry, names in $M_i^{\max}$ are exchangeable. Thus a symmetric report can be described by its size $r \in \{0, \ldots, K\}$, interpreted as a uniformly random r-subset of $M_i^{\max}$.

Fix the messages submitted by everyone other than i. Person i's message matters only when it covers every still-uncovered detected feasible name in her block. Let $C \subseteq M_i^{\max}$ be any residual pivotal set of names: if every name in C is detected and not otherwise covered, then i avoids collective punishment only if her message includes all names in C.

Compare an interior report of size $0 < r < K$ to the lottery that sends $M_i^{\max}$ with probability r/K and sends $\varnothing$ with probability $1 - r/K$. The two procedures have the same expected reporting cost, equal to ξr. For any nonempty residual pivotal set C, the endpoint lottery covers C with probability r/K. The uniform r-name report covers C with probability $\binom{K-|C|}{r-|C|}/\binom{K}{r}$, with the convention that this probability is zero if $r < |C|$. This probability is weakly below r/K for every nonempty C. Hence the endpoint lottery weakly raises the probability of covering every residual pivotal set at the same expected cost.

Therefore every interior report is weakly dominated by a lottery between silence and the maximal local report. In any undominated symmetric reporting continuation, every message used with positive probability is either $\varnothing$ or $M_i^{\max}$. □

Proof of Proposition 3.6. Fix a block B of size b and suppose that $k := |A \cap B|$ people in the block disobeyed. By Lemmas 3.4 and 3.5, it is enough to study endpoint messages inside the block.

Let α_0 be the probability that an obedient person sends the maximal local message naming all k disobedients. Let α_1 be the probability that a disobedient person sends the maximal local message naming all other disobedients.

If $k = 0$, there is no feasible name to report. If $k = b = 1$, the only person in the block disobeyed and has no feasible name to report. In both cases, everyone is silent.

Now suppose $1 \leq k < b$. We first show that disobedient people are silent. Consider an obedient person. If she sends the maximal local message, she avoids collective punishment and is not individually targeted. Her message is pivotal when the detected disobedience in the block would otherwise not be fully covered. Given the symmetric endpoint probabilities, the gross gain from reporting is

$$G_0 = (1-\alpha_0)^{b-k-1}\left[(1-\alpha_1)^k\Big(1-(1-\mu)^k\Big) + k\alpha_1(1-\alpha_1)^{k-1}\mu\right].$$

The reporting cost is ξk.

Now consider a disobedient person. Her message can help her only if no obedient person reports and no other disobedient person reports her. Conditional on that event, naming all other disobedients saves her exactly when she is not detected and at least one other disobedient is detected. Hence her gross gain from reporting is

$$G_1 = (1-\alpha_0)^{b-k}(1-\alpha_1)^{k-1}(1-\mu)\Big(1-(1-\mu)^{k-1}\Big).$$

The reporting cost is $\xi(k-1)$. If $k = 1$, a disobedient person has no feasible name to report, so she is silent. Suppose $k \geq 2$.

Write $a := 1-\alpha_0$ and $t := 1-\alpha_1$. The per-name gain for an obedient reporter is

$$\frac{G_0}{k} = a^{b-k-1}t^{k-1}\frac{t\Big(1-(1-\mu)^k\Big) + k(1-t)\mu}{k}.$$

The per-name gain for a disobedient reporter is

$$\frac{G_1}{k-1} = a^{b-k}t^{k-1}\frac{(1-\mu)\Big(1-(1-\mu)^{k-1}\Big)}{k-1}.$$

The bracketed term in G_0/k is minimized at $t = 1$, because $1-(1-\mu)^k \leq k\mu$. Hence

$$\frac{t\Big(1-(1-\mu)^k\Big) + k(1-t)\mu}{k} \geq \frac{1-(1-\mu)^k}{k}.$$

Moreover,

$$\frac{1-(1-\mu)^k}{k} > \frac{(1-\mu)\Big(1-(1-\mu)^{k-1}\Big)}{k-1}.$$

The left side is μ times the average of $1, (1-\mu), \ldots, (1-\mu)^{k-1}$, while the right side is μ times the average of $(1-\mu), \ldots, (1-\mu)^{k-1}$. Therefore, whenever $G_1 > 0$, the per-name gain from reporting is strictly larger for an obedient person than for a disobedient person.

Suppose toward a contradiction that reporting by disobedient people is active. Then $G_1 \geq \xi(k-1)$. The strict per-name comparison gives $G_0 > \xi k$, so obedient people strictly prefer to report and $\alpha_0 = 1$. But then $G_1 = 0$, contradicting active reporting by disobedient people. Hence $\alpha_1 = 0$ whenever $1 \leq k < b$.

With $\alpha_1 = 0$, the gross gain from reporting by an obedient person is $(1-\alpha_0)^{b-k-1}\big[1-(1-\mu)^k\big]$. Reporting costs ξk. Hence reporting is active if and only if $[1-(1-\mu)^k]/k > \xi$, that is, if and only if $v_k(\mu) > \xi$. If the obedient person is unique, $b-k=1$, she reports with probability one. If there are at least two obedient people, symmetric mixing requires

$$(1-\alpha_0)^{b-k-1}\big[1-(1-\mu)^k\big] = \xi k.$$

Equivalently, $\alpha_0 = 1-(\xi/v_k(\mu))^{1/(b-k-1)}$. At equality, silence is selected.

It remains to consider $k = b \geq 2$. Everyone in the block disobeyed. A disobedient person's endpoint message names the other $b-1$ disobedients. The message can help her only when she is not detected and at least one other disobedient is detected. Given the symmetric endpoint probability α_1, the gross gain from reporting is

$$G_1 = (1-\alpha_1)^{b-1}(1-\mu)\big[1-(1-\mu)^{b-1}\big].$$

The reporting cost is $\xi(b-1)$. Thus reporting is active if and only if $(1-\mu)\big[1-(1-\mu)^{b-1}\big]/(b-1) > \xi$, that is, if and only if $w_b(\mu) > \xi$. When reporting is active, symmetric mixing requires

$$(1-\alpha_1)^{b-1}(1-\mu)\big[1-(1-\mu)^{b-1}\big] = \xi(b-1).$$

Thus $\alpha_1 = 1-(\xi/w_b(\mu))^{1/(b-1)}$. At equality, silence is selected. This proves all parts of the proposition. □

A.2.3 Proofs for Section 3.2.4

Proof of Proposition 3.8. For $b = 1$, there is no peer reporting and no block liability. A person who disobeys is punished if and only if she is detected, so the unique cutoff is $x_1 = \mu$.

Now fix $b \geq 2$ and a symmetric cutoff x. Let $h \in \{0, \ldots, b-1\}$ be the number of other members of the block who disobey. Write $c_h(\mu) := 1-(1-\mu)^h$.

Suppose first that the tagged person i obeys. If $h = 0$, there is no disobedience to report and her continuation cost is zero. If $h \geq 1$, she can report all h disobedient others at cost ξh. If she remains silent, she is exposed to collective punishment when at least one of the h disobedient others is detected, an event with probability $c_h(\mu)$. Since $v_h(\mu) = c_h(\mu)/h$, the reporting condition $v_h(\mu) > \xi$ is equivalent to $\xi h < c_h(\mu)$. Hence her selected continuation cost is $O_b(h) = \min\{\xi h, c_h(\mu)\}$.

Now suppose that i disobeys. If $h \leq b-2$, the block contains $h+1$ disobedient people and at least one obedient person. By Proposition 3.6, disobedient people are silent in this state, while each obedient person reports all $h+1$ disobedient people with probability $\pi^0_{b,h+1}$. The tagged disobedient person avoids punishment only if no obedient person reports and no

disobedience in the block is detected. Therefore

$$D_b(h) = 1 - (1 - \pi^0_{b,h+1})^{b-h-1}(1-\mu)^{h+1}, \qquad h \le b-2.$$

If $h = b-1$, everyone in the block disobeys. Each other disobedient person reports all other block members with probability π^1_b. The tagged disobedient person avoids punishment only if no other person reports her and no disobedience in the block is detected. Therefore

$$D_b(b-1) = 1 - (1 - \pi^1_b)^{b-1}(1-\mu)^b.$$

Conditional on h, the payoff gain from obedience is $\Delta_b(h) := D_b(h) - O_b(h)$. Given cutoff x, each other member disobeys independently with probability $1 - F(x)$. Therefore the induced cutoff is

$$T_b(x) = \sum_{h=0}^{b-1} \binom{b-1}{h} (1-F(x))^h F(x)^{b-1-h} \Delta_b(h).$$

A symmetric cutoff equilibrium is exactly a fixed point $x = T_b(x)$.

It remains to prove existence. For each h, both $O_b(h)$ and $D_b(h)$ lie in $[0,1]$. Moreover $D_b(h) \ge O_b(h)$: whenever obedience exposes the tagged person to a reporting or silence-liability cost, disobedience exposes her to the same block event plus the possibility that she herself is detected or reported. Hence $\Delta_b(h) \in [0,1]$. The map T_b is continuous and maps $[0,1]$ into $[0,1]$. Therefore $T_b(x) - x$ is nonnegative at $x = 0$ and nonpositive at $x = 1$. By the intermediate value theorem, a fixed point exists. Since the fixed-point set is closed and nonempty, the largest symmetric cutoff equilibrium is well defined. □

A.2.4 Proofs for Section 3.2.5

Proof of Theorem 3.9. We prove the three claims in order.

First suppose $\xi \ge \mu$. Since $v_1(\mu) = \mu$, $v_k(\mu) < \mu$ for $k \ge 2$, and $w_b(\mu) < v_{b-1}(\mu)$ for $b \ge 2$, no reports are active in any state, with silence selected at equality.

Fix a block of size b. If the tagged person obeys and h other members disobey, her punishment probability is $1 - (1-\mu)^h$. If she disobeys, her punishment probability is $1 - (1-\mu)^{h+1}$. Hence the marginal value of obedience is $\mu(1-\mu)^h$. The cutoff equation is therefore

$$x = \sum_{h=0}^{b-1} \binom{b-1}{h} (1-F(x))^h F(x)^{b-1-h} \mu(1-\mu)^h = \mu\Big[1 - \mu(1-F(x))\Big]^{b-1}.$$

For $b = 1$, this gives $x_1 = \mu$. For every $b > 1$ and every $x \in [0,1]$, $F(1) < 1$ implies $1 - F(x) > 0$, so $\mu\Big[1 - \mu(1-F(x))\Big]^{b-1} < \mu$. Thus every nonsingleton block has cutoff strictly below μ. Since expected disobedience per person is $1 - F(x_b)$, the singleton partition uniquely minimizes expected disobedience. This proves part (a).

For part (b), suppose $\xi < \mu(1-\mu)$. Consider a two-person block. By Proposition 3.6, both the one-disobedient state and the all-disobedient state have active reporting. The two-

person cutoff map is $T_2(x) = 1-(\xi/\mu)\big(1-F(x)\big)$. Because $\xi < \mu(1-\mu)$, we have $\xi/\mu < 1-\mu$, and therefore $T_2(x) \geq 1 - \xi/\mu > \mu$ for every $x \in [0,1]$. Hence every two-person fixed point has cutoff strictly above $\mu = x_1$. Replacing two singleton blocks by one pair strictly lowers expected disobedience. Therefore the singleton partition is not optimal. This proves part (b).

For part (c), fix $\xi > 0$. If $\mu < \xi$, then part (a) applies, so the singleton partition is uniquely optimal for all sufficiently low μ.

It remains to show that the singleton partition is uniquely optimal for μ sufficiently close to one. Let $\rho := 1 - F(1) > 0$. Fix a nonsingleton block size $b \geq 2$ and any cutoff $x \in [0,1]$. Let H be the number of other members of the block who disobey. Since each other member disobeys with probability $1 - F(x) \geq \rho$, we have $\Pr(H \geq 1) \geq \rho$. If $H \geq 1$, an obedient person either reports at least one name at cost at least ξ or remains silent and faces collective punishment with probability at least μ. Hence $O_b(H) \geq \min\{\xi, \mu\}$. Since a disobedient person's continuation cost is at most one, $\Delta_b(H) \leq 1 - \min\{\xi, \mu\}$ on $\{H \geq 1\}$. On the complementary event, $\Delta_b(H) \leq 1$. Hence, uniformly in x and in every nonsingleton block size, $T_b(x) \leq 1 - \Pr(H \geq 1)\min\{\xi, \mu\} \leq 1 - \rho\min\{\xi, \mu\}$. For μ sufficiently close to one, the right side is strictly below μ. Thus every nonsingleton block has cutoff strictly below the singleton cutoff $x_1 = \mu$. Since there are finitely many feasible block sizes, the singleton partition is uniquely optimal for all sufficiently high μ. This proves part (c). □

A.2.5 Proofs for Section 3.2.5.2

Proof of Theorem 3.11. Fix a block size $b \geq 2$ and write $p_x := 1 - F(x)$. For every fixed nonsingleton block size, all reporting thresholds are active in every nonzero disobedience state when ξ is sufficiently small. Hence $x_b(\xi) \to 1$. Let $\rho := 1 - F(1)$.

Let $H \sim \mathrm{Bin}(b-1, p_x)$ be the number of other block members who disobey. If the tagged person obeys and $H = h$, then $O_b(h) = \xi h$ for all sufficiently small ξ, so $\mathbb{E}[O_b(H)] = \xi(b-1)p_x$.

Now suppose the tagged person disobeys. If $h \leq b-2$, let $k = h+1$ be the number of disobedients and $g = b - k$ the number of obedient reporters. If $g = 1$, the unique obedient person reports with probability one. If $g \geq 2$, then

$$(1 - \pi^0_{b,k})^g = \left(\frac{\xi}{v_k(\mu)}\right)^{\frac{g}{g-1}} = o(\xi).$$

Thus states with at least one obedient reporter contribute only $o(\xi)$ to the probability that the tagged disobedient escapes punishment.

The only first-order contribution comes from the all-disobedient state. Since $1 - \pi^1_b = (\xi/w_b(\mu))^{1/(b-1)}$, the tagged person escapes punishment with probability

$$1 - D_b(b-1) = \frac{\xi}{w_b(\mu)}(1-\mu)^b = \xi\frac{(b-1)(1-\mu)^{b-1}}{1-(1-\mu)^{b-1}}.$$

This state occurs with probability p_x^{b-1}. Since $p_{x_b(\xi)} \to \rho$, the fixed-point equation gives

$$1 - x_b(\xi) = \xi \left[(b-1)\rho + \frac{(b-1)[\rho(1-\mu)]^{b-1}}{1-(1-\mu)^{b-1}} \right] + o(\xi) = \xi A_b(\mu, \rho) + o(\xi).$$

Because F is strictly increasing, the limiting optimal nonsingleton block size minimizes $A_b(\mu, \rho)$ over integers $b \geq 2$.

Set $m := b-1$, $a := 1-\mu$, and $c_m := a^m/(1-a^m)$. Dividing the objective by the positive constant ρ, the equivalent problem is to minimize

$$H_m(a, \rho) := m + m\rho^{m-1}c_m, \qquad m \geq 1.$$

Its adjacent difference is

$$d_m(\rho) := H_{m+1} - H_m = 1 - mc_m\rho^{m-1} + (m+1)c_{m+1}\rho^m.$$

Suppose first that $m \geq 2$. The only positive critical point of d_m is $\rho_m^* := (m-1)c_m/[(m+1)c_{m+1}]$. If $\rho_m^* \leq 1$, then $d_m(\rho_m^*) = 1 - c_m(\rho_m^*)^{m-1} \geq 1 - c_m$, so $d_m(\rho) > 0$ whenever $a^m < 1/2$.

If $\rho_m^* > 1$, then d_m is decreasing on $[0,1]$ and is minimized at $\rho = 1$. Since $H_m(a,1) = m/(1-a^m)$, the sign of $d_m(1)$ is the sign of $1 - a^m(1+m\mu)$. This expression is positive because $a^m(1+m\mu) < e^{-m\mu}(1+m\mu) < 1$.

Now let $\bar{m} := \lceil 1/\mu \rceil - 1$. Then $\mu \geq 1/(\bar{m}+1)$, so

$$a^{\bar{m}} \leq \left(\frac{\bar{m}}{\bar{m}+1} \right)^{\bar{m}} \leq \frac{1}{2}.$$

The last inequality follows from $(1+1/\bar{m})^{\bar{m}} \geq 2$. It is strict except when $\bar{m} = 1$ and $\mu = 1/2$. Thus $d_m(\rho) > 0$ for every $m \geq \bar{m}$, except possibly at that boundary point. When $\mu = 1/2$ and $m = 1$, $c_1 = 1$ and $d_1(\rho) = 2c_2\rho > 0$ because $\rho > 0$. Therefore $H_{\bar{m}} < H_{\bar{m}+1} < H_{\bar{m}+2} < \cdots$, so every minimizer satisfies $m^* \leq \bar{m}$. Hence $b^*(\mu, \rho) = m^* + 1 \leq \lceil 1/\mu \rceil$.

For monotonicity, differentiate the adjacent difference with respect to a:

$$\frac{\partial d_m}{\partial a} = \rho^{m-1}a^{m-1}\left[\frac{(m+1)^2\rho a}{(1-a^{m+1})^2} - \frac{m^2}{(1-a^m)^2} \right].$$

Since $\rho \leq 1$, the bracket is nonpositive if $(m+1)\sqrt{a}(1-a^m) \leq m(1-a^{m+1})$.

To verify this inequality, set $t := \sqrt{a}$. After dividing by $1-t^2$, it is equivalent to

$$(m+1)\sum_{j=0}^{m-1} t^{2j+1} \leq m\sum_{j=0}^{m} t^{2j}.$$

Let $I := \sum_{j=1}^{m-1} t^{2j}$ and $E := 1 + t^{2m}$. The inequalities $2t^{2j+1} \leq t^{2j} + t^{2j+2}$ and $t^{2j} + t^{2(m-j)} \leq 1 + t^{2m}$ imply

$$\sum_{j=0}^{m-1} t^{2j+1} \leq I + \frac{E}{2}, \qquad I \leq \frac{m-1}{2} E.$$

Therefore

$$(m+1) \sum_{j=0}^{m-1} t^{2j+1} \leq (m+1)I + \frac{m+1}{2} E \leq mI + mE = m \sum_{j=0}^{m} t^{2j}.$$

Hence $\partial d_m / \partial a \leq 0$, so every adjacent difference is increasing in μ.

It follows that the smallest minimizer is weakly decreasing in μ. Indeed, if $\mu' > \mu$ and the smallest minimizer were larger at μ', the difference between its objective value and that of the smaller minimizer at μ would be weakly positive at μ and weakly negative at μ'. This difference is a sum of adjacent differences and is increasing in μ, a contradiction unless both values are equal at μ'. In that case the smaller value would also be a minimizer at μ', contradicting selection of the smallest minimizer. Thus $b^*(\mu, \rho)$ is weakly decreasing in μ. □

A.2.6 Proofs for Section 3.3

Proof of Lemma 3.12. Fix a block B and a signal Z_i. Let $U_i := B \setminus (\{i\} \cup Z_i)$ be the set of unobserved names.

First, reports outside B and self-reports are dominated by the same argument as in Lemma 3.4. Next, observed disobedients are exhausted before unobserved names. An observed name is known to be disobedient. An unobserved name is only possibly disobedient. Since every name has the same reporting cost, replacing an unobserved reported name with an omitted observed disobedient weakly raises the probability of covering detected disobedience and leaves the reporting cost unchanged. Hence, in any undominated symmetric reporting continuation, a report that names any unobserved person also names every observed disobedient in Z_i.

The remaining reduction is the same endpoint argument as Lemma 3.5, applied twice. Below the signal, reports are subsets of the exchangeable set Z_i. Any interior subset of Z_i is weakly dominated by a lottery between reporting no observed names and reporting all of Z_i. Above the signal, reports contain Z_i and add a subset of the exchangeable unobserved set U_i. Any interior subset of U_i is weakly dominated by a lottery between adding no unobserved names and adding all unobserved names. Expected reporting costs are unchanged, and the endpoint lotteries weakly raise the probability of covering every residual pivotal set.

Therefore every report used with positive probability is equivalent to one of the three endpoint reports $\varnothing$, Z_i, or $B \setminus \{i\}$.

If $\nu = 1$, then $Z_i = A_B \setminus \{i\}$ with probability one. Thus every name in $B \setminus (\{i\} \cup Z_i)$ is known to have obeyed. Adding such names cannot cover detected disobedience and strictly increases reporting costs when $\xi > 0$. Hence accusation is strictly dominated by evidence whenever $Z_i \neq B \setminus \{i\}$. □

Proof of Proposition 3.13. Fix $b < \infty$ and $\xi > 0$. Since punishment is bounded by one and a person can report at most $b-1$ names, the continuation payoff difference between obedience and disobedience is bounded in absolute value by $K_b := 1+\xi(b-1)$. Hence every equilibrium cutoff lies in $[-K_b, K_b]$, and the obedience probability q is bounded below by $F(-K_b) > 0$. The posterior $\pi(q,\nu) = (1-q)(1-\nu)/[q+(1-q)(1-\nu)]$ therefore converges to zero uniformly over all equilibrium obedience probabilities as $\nu \to 1$.

Fix a signal Z_i with nonempty unobserved set $U_i := B \setminus (\{i\} \cup Z_i)$. The additional benefit from moving from evidence Z_i to accusation $B \setminus \{i\}$ is bounded above by the probability that at least one unobserved name both disobeyed and is detected by the ruler. By the union bound, this probability is at most $|U_i|\mu\pi(q,\nu) \leq (b-1)\mu\pi(q,\nu)$. The additional cost of accusation is $\xi|U_i| \geq \xi$.

By uniform convergence of $\pi(q,\nu)$ to zero, there exists $\overline{\nu} < 1$ such that, for all $\nu > \overline{\nu}$ and all equilibrium obedience probabilities q, $(b-1)\mu\pi(q,\nu) < \xi$. For such ν, the benefit from adding unobserved names is strictly smaller than its cost after every signal with unobserved names. Hence accusation is not optimal after any such signal, and all reports are contained in private evidence. □

Proof of Proposition 3.14. Fix $\mu \in (0,1)$ and suppose $\xi < \mu(1-\mu)$. In the baseline model with $\nu = 1$, a two-person block has a cutoff strictly above the singleton cutoff μ by Remark 3.10.

By Proposition 3.13, for all ν sufficiently close to one, accusation is not optimal after any signal with unobserved names. Hence reports are contained in private evidence. The two-person reporting continuation and the induced punishment probabilities converge to their perfect-peer-monitoring values as $\nu \to 1$. Therefore the selected two-person cutoff converges to the baseline two-person cutoff. Since the baseline two-person cutoff is strictly above μ, there exists $\overline{\nu} < 1$ such that, for every $\nu > \overline{\nu}$, the two-person cutoff remains strictly above $x_1(\xi,\nu) = \mu$. □

Proof of Proposition 3.15. Fix $\nu \in (0,1)$ and a nonsingleton block size $b \geq 2$. For $\xi \leq 1$, every equilibrium cutoff is bounded above by $1+b$, because punishment is bounded by one and total reporting costs are bounded by $b-1$. Hence the posterior probability that an unobserved person disobeyed is bounded below by

$$\underline{\pi}_b := \frac{(1-F(1+b))(1-\nu)}{F(1+b)+(1-F(1+b))(1-\nu)} > 0.$$

Consider any reached information set at which a reporter has a nonempty unobserved set. Conditional on this information set, the probability that the unobserved set contains a detected disobedient is bounded below by a positive constant depending only on b, μ, ν, and F. If the probability that no other message covers the relevant unobserved detected disobedience were bounded away from zero along a sequence $\xi \downarrow 0$, then accusation would have a gross benefit bounded away from zero along that sequence, while its cost is at most $\xi(b-1)$. Thus, in any undominated symmetric endpoint continuation, the probability that the relevant unobserved detected disobedience remains uncovered must converge to zero as $\xi \downarrow 0$. Equivalently, cheap accusation need not make each individual accuse with probability

one, but it makes each payoff-relevant unobserved name covered by some accusation with probability approaching one.

Consequently, for every $\varepsilon > 0$, there exists $\overline{\xi} > 0$ such that, for every $\xi < \overline{\xi}$, every person in a nonsingleton block is named by some other person's message with probability at least $1 - \varepsilon$, whether she obeys or disobeys. If she obeys, other people receive no signal that clears her, so she is an unobserved name for each of them. If she disobeys, each other person either observes her, in which case evidence includes her, or fails to observe her, in which case accusation includes her with probability approaching one.

If $i \in M_{-i}$, then i is punished whether the block is collectively punished or punishment is targeted at the aggregate message. Hence, uniformly over cutoffs, $\Pr(i \in \Pi_{\mathcal{D}}(S, M) \mid a_i = 0) \geq 1 - \varepsilon$ and $\Pr(i \in \Pi_{\mathcal{D}}(S, M) \mid a_i = 1) \geq 1 - \varepsilon$ for all sufficiently small ξ. Since punishment probabilities are bounded above by one, the marginal punishment wedge between disobedience and obedience in any nonsingleton block is at most ε in the limit. Thus $\limsup_{\xi \downarrow 0} x_b(\xi, \nu) \leq 0$ for every fixed nonsingleton block size b under the selected continuation. In particular, $\limsup_{\xi \downarrow 0} x_b(\xi, \nu) \leq \mu$. Since the population is finite, there are only finitely many nonsingleton block sizes, so

$$\limsup_{\xi \downarrow 0} \max_{2 \leq b \leq n} x_b(\xi, \nu) \leq \mu.$$

Thus no nonsingleton block strictly improves on singleton punishment as reporting costs vanish. □

Proof of Corollary 3.16. Fix $\nu \in (0, 1)$. Singleton blocks give $x_1(\xi, \nu) = \mu$ for every ξ, so the singleton partition yields expected obedience $F(\mu)$ for every reporting cost.

By Proposition 3.15, as $\xi \downarrow 0$, no nonsingleton block strictly improves on singleton punishment. Since singleton blocks are always feasible,

$$\lim_{\xi \downarrow 0} Q(\xi, \nu) = F(\mu).$$

For sufficiently high reporting costs, no nonempty report is optimal: any nonempty report costs at least ξ, while the largest possible benefit from a report is avoiding one unit of punishment. With silence selected at equality, all reports are empty for all sufficiently high ξ. Coarse blocks then add liability without generating information, so singleton punishment weakly dominates every nonsingleton partition. Hence $Q(\xi, \nu) = F(\mu)$ for all sufficiently high ξ.

Now suppose that some partition $\hat{\mathcal{D}}$ strictly improves on singleton punishment at reporting cost $\hat{\xi} > 0$:

$$\frac{1}{n} \sum_{B \in \hat{\mathcal{D}}} |B| F(x_{|B|}(\hat{\xi}, \nu)) > F(\mu).$$

Then, by definition of Q, $Q(\hat{\xi}, \nu) > F(\mu)$. Combining this strict inequality with the low-cost and high-cost endpoint values gives

$$Q(\hat{\xi}, \nu) > \lim_{\xi \downarrow 0} Q(\xi, \nu) = F(\mu),$$

and $Q(\hat{\xi}, \nu)$ also exceeds the value attained when reporting costs are sufficiently high. Thus useful denunciation occurs only at intermediate reporting costs. □

A.3 Proofs for Chapter 4

A.3.1 Preliminary lemmas

The proofs below repeatedly use two facts. The first identifies maximal implementable obedience with the largest fixed point of a best-sustainable-rate map. The second is a uniform bound on binomial point masses.

Lemma A.9 (Largest fixed point)**.** *Let $T : [0,1] \to [0,1]$ be continuous with $T(1) < 1$, and let $S := \{q \in [0,1] : q \le T(q)\}$. Then $q^\dagger := \max S$ exists and satisfies $q^\dagger = T(q^\dagger)$; moreover, $q^\dagger$ is the largest fixed point of T. If, in addition, every symmetric equilibrium obedience rate q under every punishment rule satisfies $q \le T(q)$, then no punishment rule sustains obedience above $q^\dagger$.*

Proof. Since $T(0) \ge 0$, the set S is nonempty. Since T is continuous, S is closed. Hence $q^\dagger = \max S$ exists. Suppose $q^\dagger < T(q^\dagger)$. Since $T(1) < 1$, we have $1 \notin S$, so $q^\dagger < 1$. The map $q \mapsto T(q) - q$ is continuous, strictly positive at $q^\dagger$, and strictly negative at 1; by the intermediate value theorem it vanishes at some $q' \in (q^\dagger, 1)$. Then $q' \in S$ and $q' > q^\dagger$, contradicting maximality. Hence $q^\dagger = T(q^\dagger)$. Any fixed point of T lies in S, so $q^\dagger$ is the largest one.

For the final claim, if q is a symmetric equilibrium obedience rate under some punishment rule, then $q \le T(q)$ by hypothesis, so $q \in S$ and $q \le q^\dagger$. □

Lemma A.10 (Uniform local bound)**.** *For every $\varepsilon \in (0, \frac{1}{2}]$ there exists $C(\varepsilon) < \infty$ such that, for all $m \ge 1$ and all $p \in [\varepsilon, 1-\varepsilon]$,*

$$\max_{0 \le k \le m} \Pr[\mathrm{Bin}(m,p) = k] \le \frac{C(\varepsilon)}{\sqrt{m}}.$$

Proof. Write $q := 1 - p$. Stirling's double inequality, $\sqrt{2\pi n}\,(n/e)^n \le n! \le e\sqrt{n}\,(n/e)^n$ for $n \ge 1$, gives, for $1 \le k \le m-1$,

$$\binom{m}{k} p^k q^{m-k} \le \frac{e}{2\pi} \sqrt{\frac{m}{k(m-k)}} \left(\frac{mp}{k}\right)^k \left(\frac{mq}{m-k}\right)^{m-k}.$$

The last product equals $\exp\{-m\,\mathrm{KL}(k/m \,\|\, p)\} \le 1$, where KL denotes the Kullback–Leibler divergence between Bernoulli parameters. Hence

$$\binom{m}{k} p^k q^{m-k} \le \frac{e}{2\pi} \sqrt{\frac{m}{k(m-k)}} \qquad \text{for } 1 \le k \le m-1.$$

The maximum of the pmf is attained at a mode $k^* = \lfloor (m+1)p \rfloor$, which satisfies $|k^* - mp| \le 1$. If $m \ge 4/\varepsilon$, then $k^* \ge mp - 1 \ge m\varepsilon/2$ and $m - k^* \ge mq - 1 \ge m\varepsilon/2$, so $k^*(m-k^*) \ge m^2\varepsilon^2/4$

and

$$\max_k \Pr[\mathrm{Bin}(m,p)=k] \le \frac{e}{2\pi}\cdot\frac{2}{\varepsilon}\cdot\frac{1}{\sqrt{m}}.$$

For the finitely many $m < 4/\varepsilon$, the trivial bound $\max_k \Pr[\mathrm{Bin}(m,p)=k] \le 1 \le \sqrt{4/\varepsilon}/\sqrt{m}$ applies. Taking $C(\varepsilon) := \max\{e/(\pi\varepsilon),\, 2/\sqrt{\varepsilon}\}$ proves the claim. □

A.3.2 Proofs for Section 4.2

Proof of Lemma 4.1. Fix a punishment rule p. If all people other than i use cutoff t, then they obey with probability $q(t) := F(t)$, so the number of other people who disobey is $X_t \sim \mathrm{Bin}(n-1, 1-q(t))$. If i obeys, total disobedience is X_t; if she disobeys, it is $X_t + 1$. Her expected gain from obedience is $-\theta_i + \mathbb{E}[p(X_t+1) - p(X_t)] = -\theta_i + D_p(q(t))$. Thus a best response is a cutoff strategy with cutoff $b(t) := D_p(q(t))$.

Since F is continuous and $D_p(q)$ is a polynomial in q, the map b is continuous. Also $D_p(q) \in [-1,1]$ because $p \in [0,1]$, so b maps $[-1,1]$ into itself. Hence $g(t) := b(t) - t$ is continuous, with $g(-1) \ge 0$ and $g(1) \le 0$. By the intermediate value theorem, there exists $t^* \in [-1,1]$ such that $b(t^*) = t^*$. Therefore the strategy "obey if and only if $\theta_i \le t^*$" is a symmetric best response to itself. □

Lemma A.11. *Let $A(q) \in \{0,1\}^n$ have independent coordinates with $\Pr(A_i(q)=1) = 1-q$, where $A_i(q) = 1$ denotes disobedience. If $v : \{0,1\}^n \to \mathbb{R}$ is weakly decreasing in each coordinate and non-constant, then*

$$q' > q \qquad \Longrightarrow \qquad \mathbb{E}[v(A(q'))] > \mathbb{E}[v(A(q))].$$

Consequently, under the largest symmetric equilibrium selection, the ruler can solve the baseline problem by choosing a punishment rule that maximizes $q(p)$.

Proof. Let $U_1, \dots, U_n \overset{\text{iid}}{\sim} \mathcal{U}[0,1]$, and define $A_i(q) := \mathbf{1}\{U_i > q\}$. If $q' > q$, then $A(q') \le A(q)$ coordinatewise almost surely. Since v is weakly decreasing, $v(A(q')) \ge v(A(q))$ almost surely.

The inequality is strict with positive probability. Since v is weakly decreasing and non-constant, $v(0,\dots,0) > v(1,\dots,1)$. The event $\{A(q) = (1,\dots,1),\, A(q') = (0,\dots,0)\}$ has positive probability whenever $q < q'$, whether or not q or q' is at a boundary; on this event $v(A(q')) > v(A(q))$.

Thus $\mathbb{E}[v(A(q'))] > \mathbb{E}[v(A(q))]$ whenever $q' > q$. Under the selected symmetric equilibrium, any punishment rule p induces a product distribution with obedience parameter $q(p)$. Hence the ruler's expected payoff is strictly increasing in $q(p)$, and maximizing the selected obedience rate is without loss. □

Proof of Lemma 4.2. Fix $q \in [0,1]$, and let $\Lambda_q(r) := \Pr[\mathrm{Bin}(n-1, 1-q) = r]$, with $\Lambda_q(-1) = \Lambda_q(n) := 0$. Since

$$D_p(q) = \sum_{r=0}^{n-1} \Lambda_q(r)\Big(p(r+1) - p(r)\Big),$$

summation by parts gives

$$D_p(q) = \sum_{m=0}^{n} \Big(\Lambda_q(m-1) - \Lambda_q(m)\Big)p(m).$$

The binomial pmf is single-peaked, so the coefficients $\Lambda_q(m-1) - \Lambda_q(m)$ are nonpositive up to a mode and nonnegative afterward. Since the objective is linear and $p(m) \in [0,1]$, some maximizer is an upper-threshold rule $p_h(m) = \mathbf{1}\{m \geq h\}$. For this rule, $D_{p_h}(q) = \Lambda_q(h-1) = \mathrm{Piv}_h(q)$, so

$$D_p(q) \leq \max_{1 \leq h \leq n} \mathrm{Piv}_h(q) =: D^*(q) \qquad \text{for every } p, q.$$

Let $T(q) := F(D^*(q))$. Since each Piv_h is a polynomial in q and F is continuous, T is continuous, and $T(1) = F(1) < 1$. Every symmetric equilibrium obedience rate $q(p)$ satisfies $q(p) = F(D_p(q(p))) \leq T(q(p))$. By Lemma A.9, the largest fixed point $\hat{q}$ of T exists and $q(p) \leq \hat{q}$ for every rule p.

Choose any $\hat{h} \in \arg\max_{1 \leq h \leq n} \mathrm{Piv}_h(\hat{q})$. Then $\hat{q} = T(\hat{q}) = F(D^*(\hat{q})) = F(D_{p_{\hat{h}}}(\hat{q}))$, so $\hat{q}$ is a symmetric equilibrium obedience rate under $p_{\hat{h}}$. Hence $q(p_{\hat{h}}) \geq \hat{q} \geq q(p)$. □

Proof of Theorem 4.3. By Lemma 4.2 and its proof, maximal symmetric equilibrium obedience is the largest fixed point q^* of $T(q) = F(D^*(q))$, and any threshold p_h satisfying $\mathrm{Piv}_h(q^*) = D^*(q^*)$ attains it. Thus p_h is obedience-maximizing whenever $h-1$ is a mode of $\mathrm{Bin}(n-1, 1-q^*)$.

Conversely, suppose p_h is obedience-maximizing. Then its selected equilibrium is q^*, so $q^* = F(\mathrm{Piv}_h(q^*))$. If $h-1$ were not a mode, then $\mathrm{Piv}_h(q^*) < D^*(q^*)$, implying $q^* < T(q^*) = q^*$, a contradiction. Hence p_h is optimal if and only if $h-1$ is a mode of $\mathrm{Bin}(n-1, 1-q^*)$.

Finally, $q^* \in (0,1)$ because $T(0) = T(1) = F(1) \in (0,1)$. For $X \sim \mathrm{Bin}(n-1, 1-q^*)$,

$$\frac{\Pr(X = r+1)}{\Pr(X = r)} = \frac{n-1-r}{r+1} \cdot \frac{1-q^*}{q^*}.$$

This ratio is at least one if and only if $r+1 \leq n(1-q^*)$, and at most one if and only if $r+1 \geq n(1-q^*)$. Hence the modal set is $\{r \in \{0, \ldots, n-1\} : r \leq n(1-q^*) \leq r+1\}$. The strictest optimal disobedience threshold uses the smaller pivotal mode, so $h^* = n - \lfloor nq^* \rfloor$. □

Proof of Proposition 4.4. For part (a), fix n and write $T_F(q) := F(D^*(q))$. If $F_L(x) \geq F_H(x)$ for all x, then $T_{F_L}(q) \geq T_{F_H}(q)$ for every q. Hence $\{q : q \leq T_{F_H}(q)\} \subseteq \{q : q \leq T_{F_L}(q)\}$, so the largest fixed points from Lemma A.9 are ordered: $q^*(F_L, n) \geq q^*(F_H, n)$.

For part (b), fix F, and define

$$D^*_n(q) := \max_{0 \leq r \leq n-1} \Pr[\mathrm{Bin}(n-1, 1-q) = r],$$
$$D^*_{n+1}(q) := \max_{0 \leq r \leq n} \Pr[\mathrm{Bin}(n, 1-q) = r].$$

If $Y_n \sim \mathrm{Bin}(n-1, 1-q)$ and $B \sim \mathrm{Bernoulli}(1-q)$ is independent, then $Y_{n+1} := Y_n + B \sim \mathrm{Bin}(n, 1-q)$ and $\Pr(Y_{n+1} = k) = q \Pr(Y_n = k) + (1-q)\Pr(Y_n = k-1)$. Therefore

$\Pr(Y_{n+1} = k) \le D^*_n(q)$ for every k, so $D^*_{n+1}(q) \le D^*_n(q)$ for every q. Thus $T_{n+1}(q) \le T_n(q)$, and the same fixed-point comparison gives $q^*(F, n+1) \le q^*(F, n)$. □

Proof of Proposition 4.5. Under zero tolerance $p_1(r) := \mathbf{1}\{r \ge 1\}$, a person is pivotal exactly when no other person disobeys. Hence $D_{p_1}(q) = q^{n-1}$, so a symmetric equilibrium obedience rate under zero tolerance solves $q = F(q^{n-1})$. By definition, $\widehat{q}_n$ is the largest solution of this equation.

We first show that zero tolerance is obedience-maximizing if and only if $\widehat{q}_n \ge 1 - \frac{1}{n}$. If zero tolerance is obedience-maximizing, then its selected obedience rate is maximal, so $\widehat{q}_n = q^*_n$. By Theorem 4.3, zero must be a mode of $\mathrm{Bin}(n-1, 1-\widehat{q}_n)$. This is equivalent to $\widehat{q}_n \ge 1 - \frac{1}{n}$.

Conversely, suppose $\widehat{q}_n \ge 1 - \frac{1}{n}$. Then zero is a mode of $\mathrm{Bin}(n-1, 1-\widehat{q}_n)$, so $D^*(\widehat{q}_n) = \widehat{q}_n^{n-1}$ and $T(\widehat{q}_n) = F(D^*(\widehat{q}_n)) = F(\widehat{q}_n^{n-1}) = \widehat{q}_n$. Thus $\widehat{q}_n$ is a fixed point of T. There is no fixed point of T above $\widehat{q}_n$: if $q > \widehat{q}_n$, then $q > 1 - \frac{1}{n}$, so zero is a mode of $\mathrm{Bin}(n-1, 1-q)$ and $T(q) = F(q^{n-1})$; this would contradict the definition of $\widehat{q}_n$ as the largest solution of $q = F(q^{n-1})$. Hence $\widehat{q}_n$ is the largest fixed point of T, and zero tolerance is obedience-maximizing.

It remains to connect this condition to $\zeta(F)$. Let $Z_n(q) := F(q^{n-1})$ and $S_n := \{q \in [0,1] : q \le Z_n(q)\}$. By Lemma A.9, $\widehat{q}_n = \max S_n$. Since $Z_{n+1}(q) \le Z_n(q)$ for every q, one has $S_{n+1} \subseteq S_n$, so $\widehat{q}_{n+1} \le \widehat{q}_n$. Meanwhile $1 - \frac{1}{n}$ increases with n. Therefore the set of group sizes satisfying $\widehat{q}_n \ge 1 - \frac{1}{n}$ is an initial segment of the positive integers. Since $\widehat{q}_n \le F(1) < 1$ while $1 - \frac{1}{n} \to 1$, this initial segment is finite. By the definition of $\zeta(F)$, zero tolerance is obedience-maximizing if and only if $n \le \zeta(F)$. □

Proof of Lemma 4.6. The first claim is immediate when $q = 0$. Otherwise, let $B := n - O \sim \mathrm{Bin}(n, 1-q)$. Since $\lfloor nq \rfloor \le n-2$ implies $1 - q > 1/n > \log(4/3)/n$, Pinelis (2021) gives $\Pr(B > \mathbb{E}\, B) \ge \frac{1}{4}$, while $\{B > \mathbb{E}\, B\} = \{O < nq\} \subseteq \{O \le \lfloor nq \rfloor\}$. The bound is sharp as an infimum, as shown by $n = 2$ and $q \uparrow \frac{1}{2}$.

For the second claim, let (q_n) be bounded away from zero and one, and let $O_n \sim \mathrm{Bin}(n, q_n)$. Then $nq_n(1-q_n) \to \infty$, and the central limit theorem gives

$$\frac{O_n - nq_n}{\sqrt{nq_n(1-q_n)}} \Rightarrow N(0,1).$$

Since $(\lfloor nq_n \rfloor - nq_n)/\sqrt{nq_n(1-q_n)} \to 0$, it follows that $\Pr(O_n \le \lfloor nq_n \rfloor) \to \Phi(0) = \frac{1}{2}$. □

Lemma A.12. *As $n \to \infty$, $q^*_n \to q^0$ and $P^*_n \to \frac{1}{2}$.*

Proof. For each n, write $D^*_n(q) := \max_{0 \le r \le n-1} \Pr[\mathrm{Bin}(n-1, 1-q) = r]$ and $T_n(q) := F(D^*_n(q))$. Since $D^*_n(q) \le 1$, every fixed point of T_n is at most $F(1)$. Conversely, at $q^0 := F(0)$, the threshold rule that punishes only when total disobedience is n gives deterrence $(1-q^0)^{n-1}$, so $T_n(q^0) \ge F((1-q^0)^{n-1}) \ge q^0$. Since $T_n(1) = F(1) < 1$, Lemma A.9 gives a largest fixed point satisfying $q^*_n \in [q^0, F(1)] \subset (0,1)$.

Let $\underline{q} := q^0$ and $\overline{q} := F(1)$, and choose $\varepsilon \in (0, \frac{1}{2}]$ with $[1-\overline{q}, 1-\underline{q}] \subseteq [\varepsilon, 1-\varepsilon]$. By Lemma A.10,

$$\sup_{q \in [\underline{q}, \overline{q}]} D_n^*(q) \leq \frac{C(\varepsilon)}{\sqrt{n-1}} \qquad \text{for all } n \geq 2.$$

Therefore $D_n^*(q_n^*) \to 0$, and the fixed-point equation gives $q_n^* = F(D_n^*(q_n^*)) \to F(0) = q^0$.

By Theorem 4.3, punishment under the strictest optimal threshold is equivalent to $O_n \leq \lfloor nq_n^* \rfloor$ for $O_n \sim \mathrm{Bin}(n, q_n^*)$. Since (q_n^*) is bounded away from zero and one, the second part of Lemma 4.6 gives $P_n^* \to \frac{1}{2}$. □

Proof of Theorem 4.7. If $n \leq \zeta(F)$, there is nothing to prove. Suppose $n > \zeta(F)$. By Proposition 4.5, zero tolerance is not obedience-maximizing. Therefore the strictest obedience-maximizing disobedience threshold cannot be $h_n^* = 1$, so $h_n^* \geq 2$, equivalently $\lfloor nq_n^* \rfloor \leq n-2$.

Let $O_n \sim \mathrm{Bin}(n, q_n^*)$. By Theorem 4.3, punishment under the optimal threshold is equivalent to $O_n \leq \lfloor nq_n^* \rfloor$. Hence Lemma 4.6 gives $P_n^* = \Pr(O_n \leq \lfloor nq_n^* \rfloor) \geq \frac{1}{4}$. The convergence $P_n^* \to \frac{1}{2}$ follows from Lemma A.12. □

A.3.3 Proofs for Section 4.3

A.3.3.1 Proofs for Section 4.3.1

Proof of Lemma 4.8. Fix $q \in (0,1)$ and D such that $C_q(D) < +\infty$. If $D \leq 0$, the null rule is optimal and has the stated form. Suppose $D > 0$.

Let $X_q \sim \mathrm{Bin}(n-1, 1-q)$ and $B_q \sim \mathrm{Bin}(n, 1-q)$. Write $\Lambda_q(r) := \Pr(X_q = r)$, with $\Lambda_q(-1) = \Lambda_q(n) := 0$, let $\beta_q(r) := \Pr(B_q = r)$, and set $c_q(r) := \Lambda_q(r-1) - \Lambda_q(r)$. By summation by parts, the problem is

$$\min_{0 \leq p(r) \leq 1} \sum_{r=0}^{n} \beta_q(r) p(r) \qquad \text{subject to} \qquad \sum_{r=0}^{n} c_q(r) p(r) \geq D.$$

Because $\beta_q(r) > 0$ for every r, the constraint binds at any optimum: otherwise one can lower a positive coordinate slightly while preserving feasibility and strictly reducing cost. The same argument shows that an optimum assigns no punishment to a count with $c_q(r) \leq 0$.

For every r,

$$\rho_q(r) := \frac{c_q(r)}{\beta_q(r)} = \frac{r - n(1-q)}{nq(1-q)},$$

which is strictly increasing in r. Suppose an optimum has $p(r) > 0$ and $p(s) < 1$ for some $r < s$. For sufficiently small $\delta > 0$, lower $p(r)$ by $\delta/\beta_q(r)$ and raise $p(s)$ by $\delta/\beta_q(s)$. This preserves cost and changes deterrence by

$$\delta\Big(\rho_q(s) - \rho_q(r)\Big) > 0.$$

Since only counts with positive c_q are punished, one can then reduce punishment slightly and return to the binding deterrence target at strictly lower cost, a contradiction.

Thus no punished count can lie below a count that is not fully punished. If h is the lowest punished count, then $p(r) = 0$ for $r < h$, $p(r) = 1$ for $r > h$, and $p(h) = \alpha \in (0, 1]$. This is the claimed upper-threshold form. □

Proof of Lemma 4.9. Take any punishment rule p, and let $q(p)$ be its selected equilibrium. Since p provides sufficient deterrence at $q(p)$, it is feasible in the problem defining $C(q(p))$, so $C(q(p)) \leq P_p(q(p))$. Hence the ruler's payoff from p is no greater than the value of the reduced objective at $q(p)$.

Conversely, fix q such that $C(q) < +\infty$. If $q \leq q^0$, the null rule gives payoff $q^0 \geq q - \kappa C(q)$. If $q > q^0$, choose an upper-threshold minimizer p^q of $C(q)$, as permitted by Lemma 4.8. Since $F(D_{p^q}(q)) \geq q$ and $F(D_{p^q}(1)) < 1$, continuity gives a fixed point weakly above q. The selected equilibrium therefore satisfies $q(p^q) \geq q$. Expected punishment under an upper-threshold rule is nonincreasing in obedience, so $P_{p^q}(q(p^q)) \leq P_{p^q}(q) = C(q)$. Thus the payoff from p^q is at least $q - \kappa C(q)$. Taking suprema gives value equivalence.

It remains to show that the reduced problem has a solution. Targets below q^0 are dominated by q^0, so consider

$$Q := \left\{ q \in [q^0, F(1)] : D_p(q) \geq \tau(q) \text{ for some } p \in [0,1]^{n+1} \right\}.$$

The corresponding set of feasible pairs (q, p) is closed in the compact set $[q^0, F(1)] \times [0,1]^{n+1}$, so it is compact and its projection Q is compact. The minimum defining $C(q)$ is attained for every $q \in Q$.

To see that C is lower semicontinuous on Q, let $q_m \to q$ and choose minimizers p_m at q_m. Along a subsequence realizing the liminf, compactness gives $p_m \to p$. Joint continuity of $(q, p) \mapsto D_p(q)$ and $(q, p) \mapsto P_p(q)$, together with continuity of τ, implies that p is feasible at q and

$$C(q) \leq P_p(q) = \liminf_{m \to \infty} C(q_m).$$

Hence $q - \kappa C(q)$ is upper semicontinuous on Q when $\kappa > 0$ and attains a maximum. When $\kappa = 0$, the reduced objective is simply q, which also attains a maximum on Q. □

Proof of Proposition 4.10. Fix $0 \leq \kappa < \kappa'$. For each $c \in \{\kappa, \kappa'\}$, choose an optimal upper-threshold rule according to the selection in the text, and let q_c, P_c, and h_c denote its selected obedience, expected punishment, and boundary, with $h_c = n + 1$ for the null rule.

Optimality gives $q_\kappa - \kappa P_\kappa \geq q_{\kappa'} - \kappa P_{\kappa'}$ and $q_{\kappa'} - \kappa' P_{\kappa'} \geq q_\kappa - \kappa' P_\kappa$. Adding yields $P_\kappa \geq P_{\kappa'}$, and substitution into the first inequality gives $q_\kappa \geq q_{\kappa'}$.

Suppose, toward a contradiction, that $h_{\kappa'} < h_\kappa$. If $h_\kappa = n + 1$, then $P_\kappa = 0 < P_{\kappa'}$, a contradiction. Otherwise let $B_c \sim \text{Bin}(n, 1 - q_c)$. Since $q_{\kappa'} \leq q_\kappa$, $B_{\kappa'}$ first-order stochastically dominates B_κ. The rule at κ' fully punishes every count at or above h_κ, whereas the rule at κ punishes no more than that event. Hence $P_{\kappa'} \geq \Pr(B_{\kappa'} \geq h_\kappa) \geq \Pr(B_\kappa \geq h_\kappa) \geq P_\kappa$. The inequality is strict if $q_{\kappa'} < q_\kappa$; if the obedience rates are equal, it is strict because the rule at κ' also punishes a positive-probability count below h_κ. Either way this contradicts $P_{\kappa'} \leq P_\kappa$. Thus $h_{\kappa'} \geq h_\kappa$.

Let $U(c)$ be the ruler's value at cost c. Using the rule optimal at κ' as a comparison at κ gives $U(\kappa) \geq U(\kappa') + (\kappa' - \kappa) P_{\kappa'} \geq U(\kappa')$.

For welfare, write $B(q)$ for individuals' ex ante non-punishment payoff; then $B(q) = \int_q^1 F^{-1}(u)\,du$. The null rule gives the ruler q^0, so optimality implies $q_\kappa - \kappa P_\kappa \geq q^0$ and therefore $q_\kappa \geq q^0$. On $[q^0, 1]$, $B'(q) = -F^{-1}(q) \leq 0$. Since both q_κ and P_κ weakly decrease with κ, welfare $B(q_\kappa) - P_\kappa$ weakly increases. □

A.3.3.2 Proofs for Section 4.3.2

Proof of Proposition 4.11. Let $Y_{n,z} \sim \text{Bin}(n, q_{n,z})$, and write $\mu_{n,z} := nq_{n,z}$ and $\sigma_{n,z} := \sqrt{nq_{n,z}(1-q_{n,z})}$. Set $Z_{n,z} := (Y_{n,z} - \mu_{n,z})/\sigma_{n,z}$ and $a_{n,z} := (\lfloor \mu_{n,z} \rfloor - \mu_{n,z})/\sigma_{n,z}$. Because $q_{n,z} \in [\varepsilon, 1-\varepsilon]$, the Berry–Esseen bound is uniform in z, while $\sup_z |a_{n,z}| \leq [n\varepsilon(1-\varepsilon)]^{-1/2} \to 0$. Therefore

$$\sup_{z \in Z} \left| P_{n,z} - \frac{1}{2} \right| \leq \sup_{z \in Z} \left| \Pr(Z_{n,z} \leq a_{n,z}) - \Phi(a_{n,z}) \right| + \sup_{z \in Z} \left| \Phi(a_{n,z}) - \Phi(0) \right| \to 0.$$

□

A.3.3.3 Proofs for Section 4.3.3

Proof of Lemma 4.13. For a permutation $\sigma : N \to N$, write $(a_\sigma)_k := a_{\sigma^{-1}(k)}$. Fix a punishment rule p and a conjectured symmetric obedience rate q.

Fix $i, j \in N$. By Assumption 4.12(b), there exists a permutation σ with $\sigma(i) = j$ such that $w_j(a_\sigma, \theta) = w_i(a, \theta)$ for all $a \in \{0,1\}^n$ and $\theta \in \mathbb{R}$. If $b \in \{0,1\}$ and $a = (b, A^q_{-i})$, where each coordinate of A^q_{-i} equals 0 with probability q and equals 1 otherwise, then a_σ has the same distribution as (b, A^q_{-j}). Hence $\mathbb{E}[w_j(b, A^q_{-j}, \theta)] = \mathbb{E}[w_i(b, A^q_{-i}, \theta)]$ for each $b \in \{0,1\}$, and therefore

$$\mathbb{E}[w_j(0, A^q_{-j}, \theta) - w_j(1, A^q_{-j}, \theta)] = \mathbb{E}[w_i(0, A^q_{-i}, \theta) - w_i(1, A^q_{-i}, \theta)].$$

So $G_n(q, \theta)$ is well defined.

Now fix i. Conditional on a_{-i}, the gain from obedience is $w_i(0, a_{-i}, \theta_i) - w_i(1, a_{-i}, \theta_i) + p(|a_{-i}|+1) - p(|a_{-i}|)$. Taking expectations gives $G_n(q, \theta_i) + D_p(q)$. By Assumption 4.12(a), $\theta \mapsto G_n(q, \theta)$ is continuous and strictly decreasing. Since $D_p(q) \in [-1, 1]$, Assumption 4.12(c) implies that for each $(q, d) \in [0,1] \times [-1,1]$ there is a unique $t_n(q, d) \in [\underline{\theta}, \overline{\theta}]$ such that $G_n(q, t_n(q,d)) + d = 0$. Moreover, for each fixed q, the map $d \mapsto t_n(q, d)$ is strictly increasing, because $\theta \mapsto G_n(q, \theta)$ is strictly decreasing.

Hence $a_i = 0$ if and only if $\theta_i \leq t_n(q, D_p(q))$, so any symmetric equilibrium obedience rate satisfies

$$q = F\Big(t_n(q, D_p(q))\Big).$$

Assumption 4.12(d) ensures that this obedience rate is interior.

Since the action space is finite, G_n is continuous in (q, θ). Uniqueness of the root on the compact interval $[\underline{\theta}, \overline{\theta}]$ implies continuity of $(q, d) \mapsto t_n(q, d)$: if $(q_m, d_m) \to (q, d)$ and $t_n(q_m, d_m) \to t'$ along a subsequence, then continuity gives $G_n(q, t') + d = 0$, hence $t' = t_n(q, d)$. Since $D_p(q)$ is continuous in q, the map $\Psi_p(q) := F(t_n(q, D_p(q)))$ is continuous

from $[0,1]$ to $[0,1]$. Thus Ψ_p has a fixed point, and its fixed-point set is closed, hence has a largest element. □

Lemma A.13. *Fix $q \in [0,1]$. For every punishment rule p, there exists a threshold rule $p_h(m) := \mathbf{1}\{m \geq h\}$, for some $h \in \{1,\dots,n\}$, such that $D_{p_h}(q) \geq D_p(q)$. Moreover, $D_{p_h}(q) = \mathrm{Piv}_h(q)$. Hence maximal deterrence at obedience rate q is*

$$D_n^*(q) := \max_{1\leq h\leq n} \mathrm{Piv}_h(q).$$

Proof. The deterrence term $D_p(q)$ has the same expression as in the baseline, so the first paragraph of the proof of Lemma 4.2 applies verbatim: summation by parts, single-peakedness of the binomial pmf, and linearity of the objective in p yield an upper-threshold maximizer p_h with $D_{p_h}(q) = \Lambda_q(h-1) = \mathrm{Piv}_h(q)$. Taking the maximum over h gives $D_n^*(q)$. □

Proof of Proposition 4.14. For each n, define $D_n^*(q) := \max_{1\leq h\leq n} \mathrm{Piv}_h(q)$ and $T_n(q) := F(t_n(q, D_n^*(q)))$. The map T_n is continuous and satisfies $T_n(1) \leq F(\overline{\theta}) < 1$. By Lemmas 4.13 and A.13, every symmetric equilibrium obedience rate q satisfies $q \leq T_n(q)$. Conversely, if q_n^* is the largest fixed point of T_n and h maximizes $\mathrm{Piv}_h(q_n^*)$, then $q_n^* = F(t_n(q_n^*, \mathrm{Piv}_h(q_n^*)))$, so q_n^* is an equilibrium under p_h. Lemma A.9 therefore implies that q_n^* is maximal.

The modal characterization follows exactly as in Theorem 4.3. If an optimal threshold had a nonmodal pivotal count, then $\mathrm{Piv}_h(q_n^*) < D_n^*(q_n^*)$; strict monotonicity of $t_n(q,\cdot)$ and of F on the relevant interval would imply $q_n^* < T_n(q_n^*)$, contradicting the fixed-point property. Hence a threshold is optimal if and only if its pivotal count is a mode, and the strictest optimal threshold is $h_n^* = n - \lfloor nq_n^* \rfloor$.

For completeness, let $Z_n(q) := F(t_n(q, q^{n-1}))$ and let $\widehat{q}_n^Z$ be its largest fixed point, the selected zero-tolerance equilibrium. If zero tolerance is optimal, the modal characterization gives $\widehat{q}_n^Z \geq 1 - 1/n$. Conversely, suppose this inequality holds. Then zero is a mode at $\widehat{q}_n^Z$, so $T_n(\widehat{q}_n^Z) = Z_n(\widehat{q}_n^Z) = \widehat{q}_n^Z$. For every $q > \widehat{q}_n^Z$, zero remains a mode and therefore $T_n(q) = Z_n(q)$; by the definition of $\widehat{q}_n^Z$, T_n has no fixed point above it. Hence $\widehat{q}_n^Z$ is the largest fixed point of T_n, and zero tolerance is optimal.

If zero tolerance is not optimal, then $h_n^* \geq 2$, so $\lfloor nq_n^* \rfloor \leq n-2$. For $O_n \sim \mathrm{Bin}(n, q_n^*)$, punishment is equivalent to $O_n \leq \lfloor nq_n^* \rfloor$, and Lemma 4.6 gives $P_n^* \geq \frac{1}{4}$. Assumption 4.12 also implies

$$q_n^* \in [\underline{q}, \overline{q}], \qquad \underline{q} := F(\underline{\theta}) > 0, \qquad \overline{q} := F(\overline{\theta}) < 1.$$

The second part of Lemma 4.6 therefore gives $P_n^* \to \frac{1}{2}$.

Finally, if $F_L \geq F_H$ pointwise, then $T_n(q; F_L) \geq T_n(q; F_H)$ for every q, so Lemma A.9 gives $q_n^*(F_L) \geq q_n^*(F_H)$. □

Proof of Remark 4.15. Fix a punishment rule p. If the other $n-1$ people obey independently with probability q, write A_{-i}^q for their action profile, with $A_j^q = 1$ denoting disobedience, and $X_q := |A_{-i}^q| \sim \mathrm{Bin}(n-1, 1-q)$. A type-$\theta$ person's expected payoff from own action $b \in \{0,1\}$ is

$$U_\theta^b(q) := \mathbb{E}[w_i(b, A_{-i}^q, \theta)] - \mathbb{E}[p(X_q + b)],$$

and her best-response payoff is $V_\theta(q) := \max\{U_\theta^1(q), U_\theta^0(q)\}$.

Fix $q' < q''$ and couple the two profiles as in the proof of Lemma A.11: with

$$U_1, \dots, U_{n-1} \overset{\text{iid}}{\sim} \mathcal{U}[0,1] \qquad \text{and} \qquad A_j^q := \mathbf{1}\{U_j > q\} \quad \text{for} \quad q \in \{q', q''\},$$

we have $A_{-i}^{q''} \le A_{-i}^{q'}$ coordinatewise almost surely, hence also $X_{q''} \le X_{q'}$ almost surely.

By no negative externalities, $\mathbb{E}[w_i(b, A_{-i}^{q''}, \theta)] \ge \mathbb{E}[w_i(b, A_{-i}^{q'}, \theta)]$ for each b and θ. Since p is weakly increasing, $\mathbb{E}[p(X_{q''} + b)] \le \mathbb{E}[p(X_{q'} + b)]$. Therefore $U_\theta^b(q'') \ge U_\theta^b(q')$ for each b and θ, so $V_\theta(q'') \ge V_\theta(q')$.

Now let $q^L < q^H$ be two symmetric equilibrium obedience rates under the fixed rule p. At a symmetric equilibrium with obedience rate q, each type chooses a best response to that q, so her equilibrium payoff is exactly $V_\theta(q)$. Hence $V_\theta(q^H) \ge V_\theta(q^L)$ for every θ. This is weak Pareto dominance of the equilibrium at q^H over the equilibrium at q^L. □

A.3.3.4 Proofs for Section 4.3.4

Proof of Lemma 4.17. If there exist probability vectors $\pi_0, \dots, \pi_n \in \Delta(S)$ such that $\alpha(\cdot \mid a) = \pi_{|a|}$ for every $a \in \{0,1\}^N$, then for any permutation σ of N, $\alpha(\cdot \mid a^\sigma) = \pi_{|a^\sigma|} = \pi_{|a|} = \alpha(\cdot \mid a)$, since $|a^\sigma| = |a|$. So α is anonymous.

Conversely, suppose α is anonymous. For each $m \in \{0, \dots, n\}$, choose any $a^m \in \{0,1\}^N$ with $|a^m| = m$, and define $\pi_m := \alpha(\cdot \mid a^m) \in \Delta(S)$. If $a \in \{0,1\}^N$ and $m := |a|$, then a is a permutation of a^m, say $a = (a^m)^\sigma$. Hence $\alpha(\cdot \mid a) = \alpha(\cdot \mid (a^m)^\sigma) = \alpha(\cdot \mid a^m) = \pi_m = \pi_{|a|}$. Thus $\alpha(\cdot \mid a)$ depends only on $|a|$. □

Proof of Lemma 4.19. Set $\Lambda_q(m) := \Pr[\mathrm{Bin}(n-1, 1-q) = m]$, with $\Lambda_q(-1) = \Lambda_q(n) := 0$, and let $\beta_q(m) := \Pr(B_q = m)$. Summation by parts gives $d_q(j) = \sum_{m=0}^n (\Lambda_q(m-1) - \Lambda_q(m))\pi_m(s_j)$. If $\psi_q(j) > 0$, Bayes' rule and the identity $\Lambda_q(m-1) - \Lambda_q(m) = ((m - n(1-q))/(nq(1-q)))\beta_q(m)$ yield

$$\frac{d_q(j)}{\psi_q(j)} = \sum_{m=0}^n \frac{m - n(1-q)}{nq(1-q)} \Pr(B_q = m \mid s = s_j) = \frac{\mathbb{E}[B_q \mid s = s_j] - n(1-q)}{nq(1-q)}.$$

If $\psi_q(j) = 0$, then $\pi_m(s_j) = 0$ for every m, so the signal is irrelevant.

Now fix $j' > j$ with positive signal probabilities and let $\eta_{q,j}(m) := \Pr(B_q = m \mid s = s_j)$. For $m' > m$, monotonicity of the monitoring technology and Bayes' rule imply

$$\eta_{q,j'}(m')\eta_{q,j}(m) \ge \eta_{q,j}(m')\eta_{q,j'}(m).$$

Thus the posterior distribution of B_q rises with j in the MLR order. It therefore also rises in first-order stochastic dominance, so its mean weakly increases. The displayed identity proves that $d_q(j)/\psi_q(j)$ weakly increases as well. □

Proof of Remark 4.20. Fix $q \in (0,1)$. Signals with $\psi_q(j) = 0$ also have $d_q(j) = 0$ and can be ignored.

For part 1, the coefficient on p_j is $d_q(j) - \kappa\psi_q(j) = \psi_q(j)(d_q(j)/\psi_q(j) - \kappa)$. By Lemma 4.19, its sign changes at most once, from nonpositive to nonnegative, so a maximizer can be chosen upper-threshold.

For part 2, the null rule is optimal when $\overline{d} \leq 0$. When $\overline{d} > 0$, the constraint binds and no least-cost rule punishes a signal with $d_q(j) \leq 0$. The exchange argument from Lemma 4.8, with (d_q, ψ_q) in place of (c_q, β_q), shifts punishment toward weakly higher deterrence-per-cost ratios. Lemma 4.19 orders those ratios by the signal index; ties can be shifted upward without changing either cost or deterrence. Hence a least-cost rule can be chosen upper-threshold, with possible mixing at one boundary signal. □

Proof of Proposition 4.21. For each $t \in \{1, \ldots, K+1\}$, the map $q \mapsto D_t(q)$ is continuous because each $\beta_q(m)$ and $\Lambda_q(m)$ is a polynomial in q. Hence D_S^* is continuous, and so is $T(q) := F(t_n(q, D_S^*(q)))$. Moreover, $T(1) < 1$ because $t_n(1, D_S^*(1)) \leq \overline{\theta}$ and $F(\overline{\theta}) < 1$ by Assumption 4.12.

If q is a symmetric equilibrium obedience rate under some punishment rule p, then by Lemma 4.13 and Remark 4.20,

$$q = F\Big(t_n(q, D_p(q))\Big) \leq F(t_n(q, D_S^*(q))) = T(q).$$

By Lemma A.9, the largest fixed point q^* of T exists, and no punishment rule sustains obedience above q^*.

Choose any threshold t^* attaining $D_S^*(q^*)$. Then

$$q^* = F\Big(t_n(q^*, D_{t^*}(q^*))\Big),$$

so Lemma 4.13 implies that q^* is a symmetric equilibrium under p_{t^*}. Hence maximal implementable symmetric obedience is the largest fixed point of T, attained by a threshold rule. □

Proof of Remark 4.22. The stated formula is the posterior-mean identity established in the proof of Lemma 4.19. □

Proof of Remark 4.23. Delete any globally null signal. Then every remaining signal has positive probability for every $q \in (0, 1)$. The compactness and lower-semicontinuity argument in Lemma 4.9 applies verbatim in finite signal space, so optimal rules exist. Under multiplicity, maintain the largest-boundary selection used in the count model.

Fix $0 \leq \kappa < \kappa'$, and write $q := q_\kappa$, $q' := q_{\kappa'}$, $P := P_\kappa$, $P' := P_{\kappa'}$, $\ell := \ell_\kappa$, and $\ell' := \ell_{\kappa'}$. The two optimality inequalities used in Proposition 4.10 give $P \geq P'$ and $q \geq q'$.

For $r \in \{1, \ldots, K\}$, let $\Psi_r(q) := \Pr(s \geq s_r \mid q)$. Monotone likelihood ratios imply that $\Pr(s \geq s_r \mid |a| = m)$ is weakly increasing in m, while $\mathrm{Bin}(n, 1-q)$ is first-order stochastically decreasing in q. Hence $\Psi_r(q)$ is weakly decreasing in q.

Suppose $\ell' < \ell$. If $\ell = K+1$, then $P = 0 < P'$, a contradiction. Otherwise, the rule at κ' fully punishes every signal weakly above s_ℓ and also punishes a positive-probability signal below s_ℓ, whereas the rule at κ punishes no more than the event $\{s \geq s_\ell\}$. Therefore $P' > \Psi_\ell(q') \geq \Psi_\ell(q) \geq P$, again a contradiction. Thus $\ell_{\kappa'} \geq \ell_\kappa$.

Now let S be a garbling of an anonymous monitoring technology Y, so $\pi^S_m(s) = \sum_y \chi(s \mid y)\pi^Y_m(y)$. For any rule $p : S \to [0,1]$, define $\widetilde{p}(y) := \sum_s \chi(s \mid y)p(s)$. The induced expected punishment at every count is identical:

$$\bar{p}^Y(m) = \sum_y \pi^Y_m(y)\widetilde{p}(y) = \sum_s \pi^S_m(s)p(s) = \bar{p}^S(m).$$

Thus every count-contingent punishment schedule, and hence every deterrence level, attainable under S is attainable under Y. By Remark 4.20, maximal deterrence under either monotone technology is attained by a threshold rule, so $D^*_S(q) \leq D^*_Y(q)$ for every q. Consequently $T_S(q) \leq T_Y(q)$ pointwise, and Lemma A.9 gives $q^*_S \leq q^*_Y$. The same replication argument shows that the ruler's optimal payoff is weakly higher under the more informative technology. □

References

Acemoglu, D. and A. Wolitzky (2020). Sustaining cooperation: Community enforcement versus specialized enforcement. *Journal of the European Economic Association 18*(2), 1078–1122.

Alexopoulos, G. (2008). Stalin and the politics of kinship: Practices of collective punishment, 1920s–1940s. *Comparative Studies in Society and History 50*(1), 91–117.

Applebaum, A. (2003). *Gulag: A History*. Doubleday.

Baliga, S., E. Bueno de Mesquita, and A. Wolitzky (2020). Deterrence with imperfect attribution. *American Political Science Review 114*(4), 1155–1178.

Battaglini, M. (2006). Joint production in teams. *Journal of Economic Theory 130*(1), 138–167.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy 76*(2), 169–217.

Benmelech, E., C. Berrebi, and E. F. Klor (2015). Counter-suicide-terrorism: Evidence from house demolitions. *Journal of Politics 77*(1), 27–43.

Bergemann, P. (2019). *Judge Thy Neighbor: Denunciations in the Spanish Inquisition, Romanov Russia, and Nazi Germany*. Columbia University Press.

Bueno de Mesquita, B., D. P. Myatt, A. Smith, and S. A. Tyson (2024). The punisher's dilemma. *Journal of Politics 86*(2), 395–411.

Bueno de Mesquita, E. and E. S. Dickson (2007). The propaganda of the deed: Terrorism, counterterrorism, and mobilization. *American Journal of Political Science 51*(2), 364–381.

Carter, E. B. and B. L. Carter (2020). Focal moments and protests in autocracies: How pro-democracy anniversaries shape dissent in China. *Journal of Conflict Resolution 64*(10), 1796–1827.

Casper, B. and S. A. Tyson (2014). Military compliance and the efficacy of mass killings at deterring rebellion. Unpublished manuscript. Available at SSRN 2434297.

Chapkovski, P. (2021). Strike one hundred to educate one: Measuring the efficacy of collective sanctions experimentally. *PLOS ONE 16*(4), e0248599.

Chassang, S. and G. Padró i Miquel (2019). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. *The Review of Economic Studies 86*(6), 2530–2553.

Christensen, D. and F. Garfias (2018). Can you hear me now? How communication technology affects protest and repression. *Quarterly Journal of Political Science 13*(1), 89–117.

Ch'ü, T.-t. (1961). *Law and Society in Traditional China*. Mouton.

Darcy, S. (2007). *Collective Responsibility and Accountability under International Law*. Number 27 in Procedural Aspects of International Law. Transnational Publishers.

Davenport, C. (2007). State repression and political order. *Annual Review of Political Science 10*(1), 1–23.

Dickson, E. S. (2007). On the (In)Effectiveness of collective punishment: An experimental investigation. Working paper, New York University.

Duell, D., F. Mengel, E. Mohlin, and S. Weidenholzer (2024). Cooperation through collective punishment and participation. *Political Science Research and Methods 12*(3), 494–520.

Eckstein, A. M. (2005). Bellicosity and anarchy: Soldiers, warriors, and combat in antiquity. *The International History Review 27*(3), 481–497.

Egorov, G. and K. Sonin (2024). The political economics of non-democracy. *Journal of Economic Literature 62*(2), 594–636.

Esteban, J., M. Morelli, and D. Rohner (2015). Strategic mass killings. *Journal of Political Economy 123*(5), 1087–1132.

Estévez, J. L., D. Salihović, and S. V. Sgourev (2024). Endogenous dynamics of denunciation: Evidence from an inquisitorial trial. *PNAS Nexus 3*(9), pgae340.

Fairbank, J. K. (Ed.) (1978). *The Cambridge History of China, Volume 10: Late Ch'ing, 1800–1911, Part 1*. Cambridge University Press.

Fallah, B. (2024). No, punitive house demolition does not reduce suicide bombing: Revisiting evidence from the Second Intifada. *Defence and Peace Economics 35*(8), 1070–1101.

Fearon, J. D. and D. D. Laitin (1996). Explaining interethnic cooperation. *American Political Science Review 90*(4), 715–735.

Feddersen, T. and W. Pesendorfer (1997). Voting behavior and information aggregation in elections with private information. *Econometrica 65*(5), 1029–1058.

Feddersen, T. and W. Pesendorfer (1998). Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political Science Review 92*(1), 23–35.

Fitzpatrick, S. and R. Gellately (Eds.) (1997). *Accusatory Practices: Denunciation in Modern European History, 1789–1989*. University of Chicago Press.

Foucault, M. (1977). *Discipline and Punish: The Birth of the Prison*. Pantheon Books.

Garbe, L. (2024). Pulling through elections by pulling the plug: Internet disruptions and electoral violence in Uganda. *Journal of Peace Research 61*(5), 842–857.

Gellately, R. (1996). Denunciations in twentieth-century Germany: Aspects of self-policing in the Third Reich and the German Democratic Republic. *The Journal of Modern History 68*(4), 931–967.

Gieczewski, G. and K. Kocak (2025). Collective procrastination and protest cycles. *American Journal of Political Science 69*(4), 1406–1419.

Gilham, S. A. (1982). The Marines build men: Resocialization in recruit training. In R. Luhman (Ed.), *Sociological Outlook: A Text with Readings*, pp. 231–241. Wadsworth Publishing Company.

Giné, X. and D. S. Karlan (2014). Group versus individual liability: Short and long term evidence from Philippine microcredit lending groups. *Journal of Development Economics 107*, 65–83.

Goffman, E. (1958). Characteristics of total institutions. In *Symposium on Preventive and Social Psychiatry*, pp. 43–84. Walter Reed Army Institute of Research.

Gohdes, A. R. (2020). Repression technology: Internet accessibility and state violence. *American Journal of Political Science 64*(3), 488–503.

Gregory, P. R., P. J. H. Schröder, and K. Sonin (2011). Rational dictators and the killing of innocents: Data from Stalin's archives. *Journal of Comparative Economics 39*(1), 34–42.

Greif, A. (2002). Institutions and impersonal exchange: From communal to individual responsibility. *Journal of Institutional and Theoretical Economics 158*(1), 168–204.

Haggard, S. and M. Noland (2012). Economic crime and punishment in North Korea. *Political Science Quarterly 127*(4), 659–683.

Halac, M., I. Kremer, and E. Winter (2024). Monitoring teams. *American Economic Journal: Microeconomics 16*(3), 134–161.

Hassan, M., D. Mattingly, and E. R. Nugent (2022). Political control. *Annual Review of Political Science 25*, 155–174.

Hatz, S. (2019). Israeli demolition orders and Palestinian preferences for dissent. *Journal of Politics 81*(3), 1069–1074.

Heckathorn, D. D. (1988). Collective sanctions and the creation of Prisoner's Dilemma norms. *American Journal of Sociology 94*(3), 535–562.

Heckathorn, D. D. (1990). Collective sanctions and compliance norms: A formal theory of group-mediated social control. *American Sociological Review 55*(3), 366–384.

Hill, Jr., D. W. and Z. M. Jones (2014). An empirical evaluation of explanations for state repression. *American Political Science Review 108*(3), 661–687.

Holmström, B. (1982). Moral hazard in teams. *Bell Journal of Economics 13*(2), 324–340.

Hultquist, P. (2017). Is collective repression an effective counterinsurgency technique? Unpacking the cyclical relationship between repression and civil conflict. *Conflict Management and Peace Science 34*(5), 507–525.

Kalyvas, S. N. (2006). *The Logic of Violence in Civil War*. Cambridge University Press.

King, G., J. Pan, and M. E. Roberts (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review 107*(2), 326–343.

Klocker, C. (2020). *Collective Punishment and Human Rights Law: Addressing Gaps in International Law*. Routledge.

Knowles, J., N. Persico, and P. Todd (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy 109*(1), 203–232.

Kocher, M. A., T. B. Pepinsky, and S. N. Kalyvas (2011). Aerial bombing and counterinsurgency in the Vietnam War. *American Journal of Political Science 55*(2), 201–218.

Kuran, T. (1991). Now out of never: The element of surprise in the East European revolution of 1989. *World Politics 44*(1), 7–48.

Legros, P. and S. A. Matthews (1993). Efficient and nearly-efficient partnerships. *Review of Economic Studies 60*(3), 599–611.

Levinson, D. J. (2003). Collective sanctions. *Stanford Law Review 56*(2), 345–428.

Lichbach, M. I. (1987). Deterrence or escalation? The puzzle of aggregate studies of repression and dissent. *Journal of Conflict Resolution 31*(2), 266–297.

Livy (1998). *The Rise of Rome: Books One to Five*. Oxford World's Classics. Oxford University Press. Translated and edited by T. J. Luce.

Loeffel, R. (2012). *Family Punishment in Nazi Germany: Sippenhaft, Terror and Myth*. Palgrave Macmillan.

Lyall, J. (2009). Does indiscriminate violence incite insurgent attacks? Evidence from Chechnya. *Journal of Conflict Resolution 53*(3), 331–362.

McCall-Smith, K. (2016). United Nations Standard Minimum Rules for the Treatment of Prisoners (Nelson Mandela Rules). *International Legal Materials 55*(6), 1180–1205.

Montagnes, B. P. and S. Wolton (2019). Mass purges: Top-down accountability in autocracy. *American Political Science Review 113*(4), 1045–1059.

Morduch, J. (2000). The microfinance schism. *World Development 28*(4), 617–629.

Palfrey, T. R. and H. Rosenthal (1983). A strategic calculus of voting. *Public Choice 41*(1), 7–53.

Palfrey, T. R. and H. Rosenthal (1984). Participation and the provision of discrete public goods: A strategic analysis. *Journal of Public Economics 24*(2), 171–193.

Pereira, A. and J. W. van Prooijen (2018). Why we sometimes punish the innocent: The role of group entitativity in collective punishment. *PLOS ONE 13*(5), e0196852.

Persico, N. (2002). Racial profiling, fairness, and effectiveness of policing. *American Economic Review 92*(5), 1472–1497.

Pinelis, I. (2021). Best lower bound on the probability of a binomial exceeding its expectation. *Statistics & Probability Letters 179*, 109224.

Rahman, D. (2012). But who will monitor the monitor? *American Economic Review 102*(6), 2767–2797.

Rasmusen, E. (1987). Moral hazard in risk-averse teams. *RAND Journal of Economics 18*(3), 428–435.

Ritter, E. H. (2014). Policy disputes, political survival, and the onset and severity of state repression. *Journal of Conflict Resolution 58*(1), 143–168.

Ritter, E. H. and C. R. Conrad (2016). Preventing and responding to dissent: The observational challenges of explaining strategic repression. *American Political Science Review 110*(1), 85–99.

Rosendorff, B. P. and T. Sandler (2010). Suicide terrorism and the backlash effect. *Defence and Peace Economics 21*(5–6), 443–457.

Rozenas, A. (2020). A theory of demographically targeted repression. *Journal of Conflict Resolution 64*(7–8), 1254–1278.

Shadmehr, M. and D. Bernhardt (2011). Collective action with uncertain payoffs: Coordination, public signals, and punishment dilemmas. *American Political Science Review 105*(4), 829–851.

Shaw, D., H. Seaward, F. Pageau, T. Wangmo, and B. S. Elger (2023). Perceptions of collective and other unjust punishment in Swiss prisons: A qualitative exploration. *International Journal of Prisoner Health 19*(2), 241–250.

Sima Qian (1993). *Records of the Grand Historian: Qin Dynasty*. Columbia University Press. Translated by Burton Watson.

Smith, C. E. and F. Warneken (2016). Children's reasoning about distributive and retributive justice across development. *Developmental Psychology 52*(4), 613–628.

Solzhenitsyn, A. (1991). *One Day in the Life of Ivan Denisovich*. Noonday Press.

Souleimanov, E. A., D. S. Siroky, and P. Krause (2022). Kin killing: Why governments target family members in insurgency, and when it works. *Security Studies 31*(2), 183–217.

Stiglitz, J. E. (1990). Peer monitoring and credit markets. *The World Bank Economic Review 4*(3), 351–366.

Sun, J. S. (2024). The wages of repression. *Journal of Politics 86*(3), 850–863.

Taylor, M. J. (2022). Decimatio: Myth, discipline, and death in the Roman Republic. *Antichthon 56*, 105–120.

Thomas, S., C. Kelsey, and A. Vaish (2024). No one is going to recess: How children evaluate collective and targeted punishment. *Social Development 33*(2), e12730.

Ting, M. M. (2008). Whistleblowing. *American Political Science Review 102*(2), 249–267.

Tirole, J. (1996). A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *Review of Economic Studies 63*(1), 1–22.

Truex, R. (2019). Focal points, dissident calendars, and preemptive repression. *Journal of Conflict Resolution 63*(4), 1032–1052.

Tyson, S. A. (2018). The agency problem underlying repression. *Journal of Politics 80*(4), 1297–1310.

Valentino, B., P. Huth, and D. Balch-Lindsay (2004). "Draining the sea": Mass killing and guerrilla warfare. *International Organization 58*(2), 375–407.

Varian, H. R. (1990). Monitoring agents with other agents. *Journal of Institutional and Theoretical Economics 146*(1), 153–174.

Whitmeyer, J. M. (2002). The compliance you need for a cost you can afford: How to use individual and collective sanctions? *Social Science Research 31*(4), 630–652.

Wintrobe, R. (1990). The tinpot and the totalitarian: An economic theory of dictatorship. *American Political Science Review 84*(3), 849–872.

Wolton, S. and T. Yu (2026). Unmasking the enemies: A theory of denunciations. Working paper.

Yakir, P. (1972). *A Childhood in Prison*. Macmillan.

Zhou, C. (2021). Optimal size of rebellions: Trade-off between large group and maintaining secrecy. *Quarterly Journal of Political Science 16*(2), 157–183.