Strategic Abuse and Accuser Credibility

Harry Pei* Bruno Strulovici‡

March 11, 2019

Abstract: We study the interplay between the incentive for a potential abuser (*principal*) to commit crime and incentives for potential victims (*agents*) to report crime. When the punishment faced by a convicted principal is large relative to the benefit of crime, the principal's decisions to abuse agents are strategic substitutes and agents' decisions to report crime are strategic complements. This induces a negative correlation in agents' private information and a coordination motive across agents. This tension renders agents' reports arbitrarily uninformative and leads to a high probability of crime. Remedies include reducing the punishment faced by a convicted principal, and treating crimes against each individual separately.

Keywords: soft evidence, deterrence, strategic restraint, coordination, endogenous negative correlation.

JEL Codes: D82, D83, K42.

1 Introduction

Abuses of power, assaults, and other illegal activities are often hard to prove with incontrovertible evidence. To address this difficulty, judges, organizations, and the public at large often use the number of accusations leveled against an individual to assess that individual's likelihood of guilt and mitigate the risk that some accusations are driven by spite, grudges, or ulterior motives. The presumption is that the accumulation of claims against an individual makes it more likely that this individual was guilty of at least some of the claimed abuses.

This paper revisits this presumption in an environment with a rational potential abuser (*principal*) and multiple potential victims or witnesses (*agents*) whose reports are influenced by three considerations: (1) a preference for reporting the truth, (2) a concern about retaliation, and (3) possibly, a private benefit from getting the principal convicted. The crimes studied here may range from sexual and non-sexual harassment (where agents are potential victims of the principal) to financial crimes (where agents are potential whistleblowers).[1]

*Department of Economics, Northwestern University. harrydp@northwestern.edu

†Department of Economics, Northwestern University. b-strulovici@northwestern.edu

‡We thank Sandeep Baliga, Alex Frankel, George Georgiadis, Bob Gibbons, Anton Kolotinin, Matt Notowidigdo, Wojciech Olszewski, Alessandro Pavan, Larry Samuelson, Rakesh Vohra, Alex White, Alex Wolitzky, Boli Xu, and our seminar audiences for helpful comments. Strulovici acknowledges financial support from the National Science Foundation (NSF Grant No.1151410).

[1]Non-sexual harassment includes harassment based on race, religion, and gender (categorized by the EEOC as sex-based harassment of a non-sexual nature). According to the U.S. Merit Systems Protection Board (USMSPB), 26,798 harassment charges were filed with the U.S. Equal Employment Opportunity Commission (EEOC) in 2017, of which 6,696 concerned sexual harassment.

We show, perhaps paradoxically, that when the principal can expect a large punishment if a judge finds him sufficiently likely to have committed a crime, this can destroy the credibility of accusations leveled against him, and, in equilibrium, lead to a high probability of crime taking place.

In our model, the principal has several opportunities to commit a crime, each of which is associated with a distinct agent who privately observes whether the corresponding crime takes place. The principal is convicted if the ex post probability of him having committed crime, given the agents' reports, exceeds some exogenous threshold. Agents who accuse the principal incur a retaliation cost whenever their reports are insufficient to convict the principal. When considering whether to accuse the principal, an agent thus trades off the expected cost of retaliation with the benefit from punishing his abuser and, possibly, a private benefit if the principal is convicted, regardless of any actual crime committed by the principal.[2]

When the punishment faced by the principal is sufficiently large, a rational principal strategically commits few crimes in order to reduce the expected number of accusations filed against him and hence reduce the odds of being punished (Theorem 1). This strategic restraint by the principal will, in equilibrium, undermine the credibility of agents' reports and, paradoxically, leads to a high probability of crime. To see this, consider the case of two agents and suppose that two reports are required to get the principal convicted.[3] The principal's decisions about whether to abuse each agent are strategic substitutes because the principal goes unpunished if only one report is filed against him, but is punished if two reports are filed. In equilibrium, the principal abuses at most one agent. As a result, an agent who has been abused thus knows that the other agent is unlikely to file a report against the principal and hence that the abused agent's report is unlikely to result in a conviction. This weakens the abused agent's incentive to accuse the principal, since doing so incurs a high risk of retaliation.[4]

Conceptually, the principal's strategic restraint induces a *negative correlation* in agents' private information. When combined with the complementarity of agents' reporting decisions (since two reports are required for the principal to be convicted), this affects agents' reporting incentives in a way that reduces the credibility of their reports. This lack of credibility has a further perverse effect: it increases the equilibrium probability of abuse. As the punishment in case of conviction becomes arbitrarily large—or, equivalently, as the benefit

[2]The key assumptions of our model are: (1) some abuses go unreported, and (2) some charges of abuse are not deemed to have sufficient merit or credibility to lead to a conviction. Both of them are consistent with the empirical evidence on abuses and reports of abuse. According to the data released by USMSPB (2018), of the harassment charges filed in 2017, only 16% led to "*merit resolutions*," i.e., to outcomes favorable to the charging parties. Conversely, evidence provided by surveys consistently shows that a large number of assaults go unreported. For instance, a 2016 survey conducted by the USMSPB concluded that 21% of women and 8.7% of men experienced at least one of 12 categorized behaviors of sexual harassment, of which only a small fraction was followed by charges. A large scale study of sexual harassment in the U.S. military by the RAND corporation (2018) found that, depending on the military branch, between 5% and 15% of female military personal has reported being victims of sexual harassments. However, only a small fraction of which was followed by charges.

[3]We show in Theorem 1 that this property emerges endogenously for all symmetric equilibria. When there are two agents, we show that all equilibria satisfying some mild refinements *must* be symmetric. See Appendix A for details.

[4]A similar argument applies in the reverse direction: when an agent observes that he has not been abused, his posterior assigns a higher probability to the other agent having been abused compared to his prior. This encourages him to accuse the principal. Our logic goes through even when non-abused agents cannot perform such sophisticated Bayesian reasoning.

from committing crime becomes arbitrarily small—agents' reports become *arbitrarily uninformative* and the probability that the principal commits crime converges to the exogenous conviction threshold (Theorem 2). This result stands in sharp contrast with the case in which there is a single potential witness, for which we show that the agent's report becomes arbitrarily informative and the probability of abuse vanishes to zero as the punishment becomes arbitrarily large (Proposition 2).

The forces identified in this paper have another striking implication: as the punishment becomes arbitrarily large, the informativeness of individual reports becomes so weak that conviction occurs only when *all* agents accuse the principal (Proposition 7). For some intuition, suppose one report of abuse were enough to convict the principal in equilibrium. This would expose the principal to a high probability of false conviction, reflecting the high probability that at least some agents hold a grudge against the principal. This would not meet the threshold probability that is required for a conviction and hence could not be an equilibrium. Thus, more accusations must be filed in order to convict the principal. However, this requirement introduces a coordination motive among agents which, combined with the strategic restraint exerted by the principal, reduces the informativeness of individual reports. As a result, even more reports are needed, other things being equal, to reach the conviction threshold. In turn, requiring more reports for conviction renders the coordination problem even more severe across agents and reduces the informativeness of their individual reports still further, and so on. This logic implies that the principal is convicted in equilibrium only if all agents accuse him.[5]

In reality, the perceived benefits and costs of committing crime vary across individuals. This heterogeneity explains why some individuals behave as serial abusers while others exert more restraint. It also affects the belief held by an abuse victim concerning the abuser's behavior with other potential victims. To account for this heterogeneity, we consider a variation of our setting in which some abusers have *vicious* preferences in the sense that they perceive a particularly high benefit from abuse relative to the cost of punishment. We show that our results go through as long as vicious types are sufficiently unlikely relative to opportunistic types. Intuitively, an agent privately abused by a vicious principal may, consistent with his prior, erroneously believe that the principal is likely to exert some restraint with others. The negative correlation logic described earlier thus still applies from the agent's perspective, even if the principal turns out to abuse every single agent. It also explains why, when serial abusers are sufficiently rare, they may get away with committing multiple crimes.

Reciprocally, some principals may have *virtuous* preferences and never engage in crime. This possibility has little effect on our analysis. In the presence of virtuous types, an opportunistic (non-virtuous) type principal commits crime with higher probability in equilibrium. The posterior probability of guilt, conditional on all

[5]In Propositions 7 and 8, we show that even taken together, the agents' aggregated reports become arbitrarily uninformative as the punishment becomes arbitrarily large relative to the benefit of committing crime. Interestingly, the probability with which each potential victim reports *increases* with the number of potential victims. This feature stands in sharp contrast with the standard theories of public good provision, in which contributions become scarcer as the number of agents increases.

agents filing accusations, remains at the conviction threshold.

We explore several remedies to restore the informativeness of reports and reduce the probability of crime in environments with multiple agents. First, we consider an alternative conviction rule in which the judge computes, for each agent, the probability that the principal committed the crime corresponding to this agent, and then convicts the principal if the maximum of these probabilities across agents exceeds a cutoff. We show in Proposition 6 that there exists an equilibrium in which the probability of crime vanishes as the punishment becomes large. In this equilibrium, the probability of convicting the principal is linear in the number of accusations, agents' private signals are uncorrelated, and agents do not benefit from coordinating their reports. This rule restores informativeness and reduces crime. From this perspective, our analysis provides a rationale for using the *probability of each crime* when making conviction decisions, rather than the *overall probability of guilt*. However, we also show that such a rule is in general unappealing from an ex post perspective, as it may convict people whose likelihood of committing crime is lower than those that are not convicted under this rule. Imposing such a rule can thus be viewed a form of commitment imposed on the criminal justice system.

Another remedy is to restrain the number of potential victims (or witnesses) facing a given person of power, e.g., by reducing the number of direct subordinates of the principal who are vulnerable to abuse and retaliation. Indeed, we show that the equilibrium probability that the principal commits crime is increasing in the number of agents (Proposition 8). This result is driven by the lower equilibrium credibility of agents' reports as the number of agents increases, not by the higher number of opportunities to commit a crime.[6] We show that a larger number of potential victims magnifies the negative correlation effect discussed above, and reduces the informativeness of individual reports so strongly that even when all agents accuse the principal, these accusations taken together are less incriminating in equilibrium as the number of agents increases. As a result, the unconditional probability that the principal commits crime must be higher in order for the posterior probability of guilt conditional on all agents accusing the principal to reach the exogenous conviction threshold.

Next, we show that the probability of abuse *decreases* when the punishment is reduced. In particular, for some intermediate range of punishment levels, a single report suffices to convict the principal with positive probability, and in all equilibria, the conviction probabilities are concave in the number accusations. The principal's decisions to abuse different agents are now strategic complements, which induces a positive correlation in the agents' private information. The agents' incentives to coordinate their reports now lead to more credible accusations and to a lower probability of abuse. Nevertheless, this comes at a cost of increasing the number of victims conditional on abuse taking place.

Finally, we consider the use of transfers to restore the credibility of agents' reports. Intuitively, rewarding

[6]The result is thus more subtle than may initially appear. It concerns the probability that the principal commits any crime at all, an event whose feasibility does not depend on the number of potential victims he faces.

an agent who stands alone in accusing the principal offsets the retaliation cost that may be faced by the agent when the principal is not convicted. As the punishment to a convicted principal becomes arbitrarily large, such rewards can restore the informativeness of the agents' reports to a level that becomes arbitrarily close to that of the single-agent benchmark (Proposition 5). However, these transfers are not budget-balanced. They are costly to the social planner and create incentives for the principal and the agents to collude.

Motivated by the aforementioned drawbacks, we consider budget-balanced transfer schemes. By construction, these schemes require negative transfers to some of the agents, which are hard to implement in some applications. But even if one could ignore this issue, we show that budget-balanced transfer schemes are of limited value for restoring the credibility of agents' reports. Intuitively, these schemes must punish agents who fail to accuse the principal when others do, which induce an additional coordination motive among the agents. Conditional on an agent accusing the principal, it becomes more beneficial for other agents to also accuse the principal in order to avoid the negative transfer. Balanced-budget transfers thus remove the coordination motives arising from retaliation costs while generating a new coordination motive, and this explains their limited effectiveness.

The insights from our baseline model can be extended in several directions. First, the principal could face a larger punishment if the probability that he is believed to have committed multiple abuses is high enough. Alternatively, the principal could face decreasing marginal returns from committing multiple crimes. These departures only strengthen the principal's incentive to abuse few agents, and consequently, the induced negative correlation in the agents' private information. Second, the retaliation cost suffered by each reporting agent may be decreasing in the number of reporting agents. This feature strengthens the agents' coordination motive, which exacerbates the effects described in the baseline model. Third, the principal may hold private information concerning the number of potential victims or witnesses. In this case, we show that the principal's incentive to commit crime is stronger when this number is smaller. Intuitively, the expected number of reports is smaller when there are fewer potential victims, other things being equal. The negative correlation emphasized throughout this paper arises in this setting as well, because an agent who has been abused infers that the set of potential victims is small and expects few reports to be filed by other agents. Finally, the insights of our baseline model remain valid when agents directly care about crimes committed against (or observed by) other agents. The endogenous negative correlation emphasized earlier has an even more pronounced effect under this variation. This is because aside from coordination motives, each agent's payoff directly depends on other agents' private information. As a result, his reporting decision attaches more weight to his belief about other agents' signals, and hence is relatively less responsive to his private information.

Related Literature This paper contributes to several strands of literature studying information aggregation and collective decision making, games of coordination, law and economics, and the elicitation of private

information, which plays a major role in mechanism design.

First, we identify a novel economic mechanism to explain the failure of information aggregation. Existing models of social learning attribute such failures to informational externalities (Banerjee 1992, Bikhchandani, Hirshleifer, and Welch 1992, Smith and Sørensen 2000) or payoff externalities (Scharfstein and Stein 1990, Ottaviani and Sørensen 2000). In these papers, agents fail to act on their private information only when they observe informative actions taken by their predecessors. By contrast, agents cannot observe one another's actions in our model, and an increase in the number of agents can result in agents' reports becoming *less* informative, even when taken collectively. In social learning models, agents' signals are either conditionally independent or positively correlated. In this paper, by contrast, agents' private information (i.e., whether they have observed a crime or not) is negatively correlated. It is the combination of this negative correlation and agents' incentives to coordinate that undermines the informativeness of their aggregate reports.[7]

Failures of information aggregation also arise in models of voting and, more generally, of collective decision making. They can be caused by individual biases (Morgan and Stocken 2008) or by voters' believed probabilities for which they are pivotal (Austen-Smith and Banks 1996, Bhattacharya 2013). When voters' payoffs from a reform are negatively correlated, Ali, Mihm, and Siga (2018) show that collective decisions are socially inefficient under supermajority rules. Intuitively, if a voter is pivotal, then a sizable share of other voters is in favor of the reform. This, together with the assumed negative correlation of payoffs across voters, implies that the pivotal voter is unlikely to benefit from the reform himself. In the present paper, the negative correlation is endogenous and concerns agents' private information, rather than their payoffs. The failure of information aggregation stems from the interaction between this negative correlation and the agents' coordination motives. Another distinctive feature of our model is that the voting rule and the correlation structure of the agents' private information are both endogenous. This distinguishes our model from earlier models of strategic voting, such as Feddersen and Pesendorfer (1996), in which these elements are exogenous.

Second, the coordination motive that arises endogenously in our model is reminiscent of the extensive literature on global games. In Carlson and Van Damme (1993) and Morris and Shin (1998), agents receive conditionally independent private signals about a common state of the world. In Baliga and Sjöström (2004) and Chassang and Padró i Miquel (2010), each agent privately observes his value for a decision, and the values are independent across agents. In contrast, the agents in the present paper have *negatively correlated* private signals. This combined with their coordination motives, causes their decisions to be largely uninformative about their private information.

Third, our paper contributes to the law and economics literature by endogenizing the credibility of a

[7] Strulovici (2018) studies a sequential learning model in which an agent is less likely to have an informative signal, other things being equal, if another agent has found such a signal. This *information attrition* may be viewed as a form of negative correlation across agents' signals, which also has adverse effects on learning.

particular type of evidence—witness testimony—and studying the interplay between an individual's incentive to commit crimes and potential victims' incentives to report crimes. Our conclusion that a lower punishment can reduce the occurrence of crime stands in contrast to Becker's (1968) well-known observation that maximal punishments save on law-enforcement costs, and to the optimal judicial mechanisms of Siegel and Strulovici (2018), who find that extreme punishments are optimal in a binary-type environment.[8] Our conclusions thus highlight the subtlety of designing appropriate punishments for crimes that are nonverifiable, and for which judicial decisions are based on reports whose credibility depends on accusers' beliefs and incentives.[9]

Our model of strategic abusers differs from the recent work of Lee and Suen (2018), in which a criminal commits a crime against each of the two agents with some fixed exogenous probability, and the focus is on the timing of reports by victims and by false accusers. They provide an explanation for the well-documented fact that victims sometimes delay their accusations. Their analysis and ours consider complementary aspects of the victims' reporting incentives. In particular, the potential criminal's strategic restraint that emerges endogenously in our model, and the negative correlation that it induces on the agents' private information, are distinctive features of our analysis.

Lastly, the failure to elicit correlated private information from multiple informed agents stands in contrast to the results in Crémer and McLean (1985, 1988), which provide a convex independence condition under which such information can be elicited via a budget-balanced mechanism. In our model, budget-balanced transfers cannot be used to restore full informativeness. This is because agents' types are two-dimensional: one dimension concerns whether the agent was abused, and another dimension concerns the agent's private benefit from accusing the principal. This second dimension causes the convex independence condition to fail, because agents with the same abuse status but different private benefits hold the same belief about the others' types.

The rest of this paper is organized as follows. We introduce our baseline model in section 2 and state our main results in section 3. Section 4 examines two solutions to restore informativeness. Section 5 studies extensions of the baseline model. Section 6 concludes by discussing the interpretations and applicability of our results, the trade-offs that our analysis reveals on the design of laws and institutions, and the robustness of our insights in other settings.

[8] Stigler (1970) observes that several punishment levels should optimally be used when criminals can choose between different levels of crime. The rationale is to provide marginal incentives not to commit the worst crimes. In this scenario, applying the maximal punishment to the worst crimes remains optimal from the perspective of deterring crimes.

[9] In settings that contrast with ours, Silva (2018) and Baliga, Bueno de Mesquita, and Wolitzky (2019) consider models with multiple potential suspects and at most one crime. Silva (2018) constructs a scheme that elicits truthful confessions among suspects. In Baliga, Bueno de Mesquita, and Wolitzky (2019), one of the potential assailants has an opportunity to commit crime. In both papers, the negative correlation in suspects' types is exogenous.

2 Baseline Model

Consider the following three-stage game between a principal (e.g., the manager of a firm), n agents (e.g., the principal's subordinates) and an evaluator (e.g., a judge, or the board of trustees). In stage 1, the principal chooses an n-dimensional vector $\boldsymbol{\theta} \equiv (\theta_1, ..., \theta_n) \in \{0,1\}^n$, in which $\theta_i = 0$ means that the principal abuses agent i, and $\theta_i = 1$ means that the principal abstains from doing so. In stage 2, agent $i \in \{1, 2, ..., n\}$ privately observes: (1) the principal's choice $\theta_i \in \{0,1\}$; and (2) the realization of a private shock $\omega_i \in \mathbb{R}$; before deciding whether to file a report against the principal or not. Let $a_i \in \{0,1\}$ denote agent i's decision, with $a_i = 1$ if he reports and $a_i = 0$ otherwise. We interpret ω_i as a utility shock that affects agent i's preference towards the principal. We assume that the variables $\omega_1, \omega_2, ..., \omega_n$ are independently and normally distributed with mean $\mu \geq 0$ and variance σ^2.[10] We use $\Phi(\cdot)$ and $\phi(\cdot)$ to denote these variables' common cdf and pdf.

For technical purposes, we allow each agent to be mechanical with some small probability $1 - \delta$ and strategic with probability $\delta \in (0,1)$. A strategic agent chooses whether to report or not in order to maximize his payoff, while a mechanical agent files a report with some exogenous probability $\alpha \in (0,1)$. Whether an agent is strategic or mechanical is independent of $\{\omega_1, ..., \omega_n\}$ and of whether other agents are strategic or mechanical. The primary focus of our analysis is on the strategic agents. Allowing agents to be non-strategic serves a technical purpose. It can also be viewed as a form of robustness check in case some agents do not respond to incentives.[11] We are primarily interested in settings in which mechanical types are arbitrarily rare, although our characterization results do not rely on this.

In stage 3, the evaluator observes the vector of reports $\boldsymbol{a} \equiv (a_1, ..., a_n) \in \{0,1\}^n$ and updates his belief about $\Pi_{i=1}^n \theta_i$, i.e., whether the principal is *guilty* (in which case $\Pi_{i=1}^n \theta_i = 0$) or *innocent* (in which case $\Pi_{i=1}^n \theta_i = 1$). He then chooses whether to convict the principal. We denote this decision by $s \in \{0,1\}$, with $s = 0$ if the principal is convicted and $s = 1$ if the principal is acquitted.

Payoffs: The evaluator's payoff function is quadratic, given by:

$$-\left(s - (\pi^* - \frac{1}{2}) - \Pi_{i=1}^n \theta_i\right)^2, \tag{2.1}$$

where $\pi^* \in (0,1)$ is an exogenous parameter. The optimal choice of $s \in \{0,1\}$ is the one that is closest to the evaluator's bliss point: $\pi^* - \frac{1}{2} + \mathbb{E}\left[\Pi_{i=1}^n \theta_i \middle| \boldsymbol{a}\right]$, which increases with the posterior probability that the principal

[10] Our analysis applies to any distribution that has full support and a thin left tail. The assumption that $\mu \geq 0$ is needed for the comparative statics (statement 4 of Theorem 1 and Proposition 8), but is not required for other results.

[11] As will be explained in section 3, having a positive fraction of non-strategic agents, no matter how small, (1) rules out "bad" trivial equilibria, and (2) guarantees the existence of equilibrium that survives our refinements. Our insights are robust under alternative specifications of the mechanical types' strategies. See section 5 and Online Appendix H.2.

is innocent. Therefore, the evaluator's optimal strategy takes the following cutoff form:

$$s\begin{cases} = 0 & \text{if} \quad \Pr\left(\Pi_{i=1}^{n}\theta_i = 0 \middle| \boldsymbol{a}\right) > \pi^* \\ \in \{0,1\} & \text{if} \quad \Pr\left(\Pi_{i=1}^{n}\theta_i = 0 \middle| \boldsymbol{a}\right) = \pi^* \\ = 1 & \text{if} \quad \Pr\left(\Pi_{i=1}^{n}\theta_i = 0 \middle| \boldsymbol{a}\right) < \pi^*, \end{cases} \tag{2.2}$$

that is, the principal is convicted only when the probability with which he is guilty is at least π^*. The principal's payoff is:

$$\sum_{i=1}^{n}(1-\theta_i) - L(1-s), \tag{2.3}$$

in which $L > 0$. According to (2.3), the principal's marginal benefit from committing each crime is normalized to 1, and L is the principal's loss from conviction relative to his benefit of committing crimes. For each strategic agent, his payoff when the principal is convicted is normalized to 0. When the principal is acquitted, his payoff is:

$$\omega_i + b\theta_i - ca_i \tag{2.4}$$

where $b > 0$ is his preference for convicting his abuser, and $c > 0$ is his loss from the principal's retaliation, which is incurred when he accuses the principal but fails to convict him.

Interpretation of Payoffs: In our model, the principal strictly benefits from committing more crimes and is punished if he is believed to be guilty with high enough probability. This induces a trade-off for the principal between the benefits of crime and the risk of punishment. Our main interest concerns the case in which L is large enough. This includes situations in which conviction entails the loss of a lucrative position and a promising career (e.g., when the principal is a manager of a firm), significant harm to the principal's public image (e.g., if the principal is a public figure), a loss of power (e.g., when the principal is a politician or bureaucrat), or being ostracized from one's community.

An agent's relative payoff when the principal is acquitted (i.e., $s = 1$) rather than convicted (i.e., $s = 0$) depends on three terms. First, it is affected by an idiosyncratic taste towards the principal, modeled by the i.i.d. preference shock ω_i. Second, each agent is more inclined to have the principal convicted when he has been abused, as captured by the parameter $b > 0$.[12] Since (2.4) is agent i's utility at the reporting stage *after* the abuse has taken place, b does not capture the direct utility loss from the abuse, which has been sunk. Rather, it may capture an agent's preference for justice, or his disutility from continuing to interact with a principal who

[12] A more general way of modeling this feature is to assume that the distribution of ω_i is increasing in θ_i in the sense of MLRP: an agent's preference towards acquitting the principal is stronger when the principal has not abused him. Our formulation is a particular case of this, since the $b\theta_i$ is equivalent to a mean shift of ω_i from μ to $\mu + b\theta_i$.

has abused him in the past, such as the increased risk of being abused again in the future.

Third, filing a report costs an agent c if the principal is acquitted. When the principal is an influential person or when the agents are the principal's subordinates, c may capture the principal's retaliation against an accuser, which is stronger when the principal remains in power. In other applications, c may capture the social stigma suffered by the reporting agent or his monetary cost of going through the judicial process (such as paying the lawyer fee). More generally, our results apply as long as each agent's reporting cost is *strictly higher* when the principal is acquitted.[13]

Since agents can accuse the principal regardless of whether they have been abused or not, the evaluator must take into account this credibility problem when deciding whether to convict the principal. The cutoff probability π^* measures the evaluator's attitude towards the trade-off between convicting an innocent defendant and acquitting a guilty one. We take π^* as an exogenous parameter to reflect the social preferences over this tradeoff, instead of as a design instrument.

Depending on the application, the evaluator could be a judge when abuses are of a criminal nature and the punishment is enforced by the criminal justice system. The evaluator could also be the owner of a firm, of which the principal is a manager whose misbehavior can hurt the firm's public image. When the principal's behavior is legal but immoral, the evaluator stand for other members of the principal's community, in which case the punishment takes the form of ostracism or the loss of status within the community.

3 Results

Our analysis focuses on Bayesian Nash equilibria that satisfy two refinements: *presumption of innocence* and *symmetry*, defined in section 3.1. We first derive and compare the equilibrium outcomes with a single agent and with two agents, showing that large differences already emerge between these two cases. We generalize our analysis to three or more agents in section 5.1 and derive comparative statics with respect to the number of agents. Although we focus on symmetric equilibria in our analysis, we show in Appendix A that symmetry emerges endogenously for the case of two agents under some natural and arguably weak refinements.[14]

[13] In section 5, we show that our results continue to hold when: agents directly care about crimes committed against other agents; the effect of retaliation against a reporting agent is decreasing in the number of reporting agents; the principal faces a larger punishment than L when he is believed to have committed multiple crimes; the principal has decreasing marginal returns from committing crimes; and the principal's utility from abusing agents is privately known and can be arbitrarily small or even negative.

[14] In Appendix A, we show that when there are two agents, all sequential equilibria are symmetric under a monotonicity refinement (Axiom 2), which requires that the principal be convicted with a weakly higher probability when a larger subset of agents report against him; and a "properness" refinement (Axiom 3) that resembles Myerson (1978)'s definition of proper equilibrium, which requires that at each off-path history, each agent believes that the principal is arbitrarily less likely to make more costly mistakes.

3.1 Equilibrium Refinement & Preliminary Analysis

A Bayesian Nash equilibrium consists of $\left\{(\sigma_i)_{i=1}^n, \boldsymbol{\pi}, q\right\}$ in which $\sigma_i : \mathbb{R} \times \{0,1\} \to \Delta\{0,1\}$ is agent i's strategy (when the agent is strategic), $\boldsymbol{\pi} \in \Delta(\{0,1\}^n)$ is the principal's strategy, and $q : \{0,1\}^n \to [0,1]$ is the mapping from the vector of reports to the conviction probabilities. Our refinement introduces a regularity condition on the mapping q:

Axiom 1 (Presumption of Innocence)**.** $q(0,0,...,0) = 0$.

Axiom 1 requires that the principal is not convicted when nobody accuses him. It rules out equilibria in which the principal abuses all agents and is convicted no matter what. Those equilibria are not only unrealistic, but are also against the idea that a defendant should be convicted based on the accusations filed against him rather than on the sole basis of a prior belief. Lemma 3.1 shows that under presumption of innocence, the ex ante probability of crime is interior and the conviction probabilities are responsive to every agent's report.

Lemma 3.1. *In every Bayesian Nash equilibrium that satisfies presumption of innocence,*

1. $\Pi_{i=1}^n \theta_i = 0$ *occurs with probability strictly between* 0 *and* 1.[15]

2. *For every* $i \in \{1,...,n\}$*, there exists* $a_{-i} \in \{0,1\}^{n-1}$ *such that* $q(1,a_{-i}) \neq q(0,a_{-i})$.

Proof. For statement 1, suppose towards a contradiction that $(\theta_1,...,\theta_n) = (1,1,...,1)$ occurs with probability 1. Since $\alpha \in (0,1)$, every $\boldsymbol{a} \in \{0,1\}^n$ occurs with strictly positive probability. Given that the prior probability of $\Pi_{i=1}^n \theta_i = 0$ is 0, the posterior probability of $\Pi_{i=1}^n \theta_i = 0$ under every reporting profile is also 0. Therefore, the principal is acquitted for sure under every reporting profile, which gives him a strict incentive to abuse all agents. This leads to a contradiction. Next, suppose towards a contradiction that $\Pi_{i=1}^n \theta_i = 0$ occurs with probability 1. Then $\alpha_1 \in (0,1)$ implies that $\Pr(\Pi_{i=1}^n \theta_i = 0 | \boldsymbol{a}) = 1$ for every $\boldsymbol{a} \in \{0,1\}^n$. As a result, $q(0,...,0) = 1$. This contradicts Axiom 1.

For statement 2, suppose towards a contradiction that there exists an agent i, such that $q(1,a_{-i}) = q(0,a_{-i})$ for all $a_{-i} \in \{0,1\}^{n-1}$. Then the principal's marginal cost of abusing agent i is 0. As a result, the principal has a strict incentive to set $\theta_i = 0$. This means that the prior probability with which $\Pi_{i=1}^n \theta_i = 0$ is 1 and the principal is convicted under every $\boldsymbol{a}$. This contradicts Axiom 1. □

We say that an equilibrium is symmetric if all agents adopt the same strategy (i.e., $\sigma_1 = \sigma_2 = ... = \sigma_n$), and the principal's strategy $\boldsymbol{\pi}$ treats the agents symmetrically. Axiom 1 leads to a monotonicity property of the equilibrium conviction probabilities when focusing attention on symmetric equilibria:

[15] This is related to albeit different from a well-known result in the literature on inspection games (Dresher 1962), in which crime occurs with positive probability due to the inspector's cost of inspection. In our model, it is because the principal is convicted only when the posterior probability of crime exceeds an interior cutoff.

Lemma 3.2 (Monotonicity)**.** *In every symmetric Bayesian Nash equilibrium that satisfies presumption of innocence, for every* $\boldsymbol{a}, \boldsymbol{a'} \in \{0,1\}^n$ *with* $\boldsymbol{a} \succ \boldsymbol{a'}$*, we have* $q(\boldsymbol{a}) \geq q(\boldsymbol{a'})$*.*

Proof. We begin by showing that $\frac{\Pr(a_i=1|\theta_i=0)}{\Pr(a_i=1|\theta_i=1)} > 1$ for all i. When the equilibrium is symmetric, the value of this ratio is the same for all agents. First, suppose that the ratio is to equals 1. Then, the principal's marginal cost of abusing each agent is 0, which provides him a strict incentive to abuse all agents, which leads to a contradiction. Next, suppose the ratio is strictly less than 1. According to Lemma 3.1, the symmetry of the principal's equilibrium strategy implies that each agent is abused with strictly positive probability. As a result, for every $\boldsymbol{a} \neq (0,0...,0)$, we have:

$$\Pr(\Pi_{i=1}^n \theta_i = 0|\boldsymbol{a}) < \underbrace{\Pr(\Pi_{i=1}^n \theta_i = 0|(0,0,...,0)) \leq \pi^*}_{\text{since } q(0,...,0)=0}.$$

The above inequality implies that $q(\boldsymbol{a}) = 0$ for all $\boldsymbol{a} \in \{0,1\}^n$, which gives the principal a strict incentive to abuse all agents. This leads to a contradiction. Since $\frac{\Pr(a_i=1|\theta_i=0)}{\Pr(a_i=0|\theta_i=1)} > 1$ for all i and each agent is abused with positive probability, this implies that for every $\boldsymbol{a}, \boldsymbol{a'} \in \{0,1\}^n$ with $\boldsymbol{a} \succ \boldsymbol{a'}$, we have $\Pr(\Pi_{i=1}^n \theta_i = 0|\boldsymbol{a}) > \Pr(\Pi_{i=1}^n \theta_i = 0|\boldsymbol{a'})$. As a result, $q(\boldsymbol{a}) \geq q(\boldsymbol{a'})$. □

Lemma 3.2 implies that in every symmetric equilibrium that satisfies presumption of innocence, each agent's report is a move against the principal, which leads to a weak increase in the conviction probability, as well as a strict increase in the conviction probability in some circumstances. As a result, every agent's equilibrium strategy is characterized by two cutoffs: ω^* and ω^{**}, such that agent i reports either when $\omega_i \leq \omega^*$ and $\theta_i = 0$, or when $\omega_i \leq \omega^{**}$ and $\theta_i = 1$.

3.2 Single-Agent Benchmark

If there is only one agent, then according to Axiom 1, the principal is acquitted unless the agent sends a report. Let q_s denote the probability that the principal is convicted if the agent sends a report. If the agent is abused $(\theta = 0)$, then he prefers to send a report when

$$\omega \leq (1-q_s)(\omega - c), \quad \text{or equivalently,} \quad \omega \leq \omega_s^* \equiv -c\frac{1-q_s}{q_s}. \tag{3.1}$$

Similarly, if the agent is not abused $(\theta = 1)$, then he prefers to send a report when

$$\omega + b \leq (1-q_s)(\omega + b - c), \quad \text{or equivalently,} \quad \omega \leq \omega_s^{**} \equiv -b - c\frac{1-q_s}{q_s}. \tag{3.2}$$

Subtracting (3.1) from (3.2), this shows that $\omega_s^* - \omega_s^{**} = b$. The *informativeness* of the agent's report is measured by the likelihood ratio that the agent sends a report as a function of whether he has observed a crime:

$$\mathcal{I}_s \equiv \frac{\Pr(\text{agent reports} \mid \theta = 0)}{\Pr(\text{agent reports} \mid \theta = 1)} = \frac{\delta\Phi(\omega_s^*) + (1-\delta)\alpha}{\delta\Phi(\omega_s^{**}) + (1-\delta)\alpha}. \tag{3.3}$$

The agent's report is thus completely uninformative if $\mathcal{I}_s = 1$ and perfectly informative if $\mathcal{I}_s = +\infty$.

As the probability δ that the agent is strategic goes to 1, this informativeness ratio converges to $\Phi(\omega_s^*)/\Phi(\omega_s^{**})$. Let π_s denote the probability that the principal commits crime in equilibrium. We provide conditions on the existence and uniqueness of equilibrium and offer a characterization:

Proposition 1. *There exists an equilibrium that satisfies Axiom 1 if and only if*

$$\delta L\Big(\Phi(0) - \Phi(-b)\Big) \geq 1. \tag{3.4}$$

When (3.4) holds as a strictly inequality, this equilibrium is unique and characterized by a tuple $(\omega_s^*, \omega_s^{**}, q_s, \pi_s) \in \mathbb{R} \times \mathbb{R} \times (0,1] \times [0,1]$ *that satisfies (3.1), (3.2),*

$$\delta q_s\Big(\Phi(\omega_s^*) - \Phi(\omega_s^{**})\Big) = 1/L \tag{3.5}$$

and

$$\mathcal{I}_s \frac{\pi_s}{1-\pi_s} = \frac{\pi^*}{1-\pi^*}. \tag{3.6}$$

Next, we study the limiting properties of this unique equilibrium when the punishment becomes arbitrarily large and the fraction of mechanical type agent becomes arbitrarily small.

Proposition 2. *As* $L \to \infty$*, we have* $q_s \to 0$*,* $\omega_s^*, \omega_s^{**} \to -\infty$*. Moreover,* $\lim_{L\to\infty}\lim_{\delta\to 1}\mathcal{I}_s = \infty$ *and* $\lim_{L\to\infty}\lim_{\delta\to 1}\pi_s = 0$*.*

The proofs of Propositions 1 and 2 are in Appendix C. Proposition 2 shows that, as $\delta \to 1$ and $L \to \infty$ in the stated order, the agent's report becomes arbitrarily informative and the equilibrium probability of crime vanishes to 0. More generally, when δ is large enough, increasing L both improves the informativeness of the agent's report and decreases the equilibrium probability of crime.[16]

[16]Formally, for every L' and L'' such that $L' > L''$, there exists $\overline{\delta} \in (0,1)$ such that when $\delta > \overline{\delta}$ and (3.4) holds with $\overline{\delta}$ and L'', $\mathcal{I}_s$ is strictly higher and π_s is strictly lower when $L = L'$ than when $L = L''$.

3.3 Two Agents

We now study the properties of all symmetric Bayesian Nash equilibria that satisfy presumption of innocence (or *equilibrium* for short) when there are two agents and compare it to the single-agent benchmark. We start by establishing the existence of such equilibria:

Proposition 3. *If L is large enough, then there exists an equilibrium.*

The proof, in Online Appendix A, is based on Brouwer's fixed point theorem. It generalizes to any arbitrary number of agents as well as to alternative specifications of the mechanical types' strategies. In what follows, we focus on values of L for which an equilibrium exists. Our first theorem characterizes the common properties of *all* equilibria when L is large and compares them to the unique equilibrium in the single-agent benchmark.

Theorem 1. *There exists $\overline{L} \in \mathbb{R}_+$ such that when $L > \overline{L}$, every equilibrium is characterized by a tuple $(\omega_m^*, \omega_m^{**}, q_m, \pi_m) \in \mathbb{R}_- \times \mathbb{R}_- \times [0,1] \times [0,1]$ such that*

1. *For every $i \in \{1,2\}$, agent i reports when $\{\omega_i \leq \omega_m^*$ and $\theta_i = 0\}$ or $\{\omega_i \leq \omega_m^{**}$ and $\theta_i = 1\}$.*
2. *The principal chooses $(\theta_1, \theta_2) = (1,1)$ with probability $1 - \pi_m$, $(\theta_1, \theta_2) = (1,0)$ with probability $\pi_m/2$, and $(\theta_1, \theta_2) = (0,1)$ with probability $\pi_m/2$.*
3. *The conviction probabilities satisfy $q(0,0) = q(0,1) = q(1,0) = 0$ and $q(1,1) = q_m$.*
4. *Compared to the unique equilibrium in the single-agent benchmark, we have: $\omega_m^* > \omega_s^*$, $\omega_m^{**} > \omega_s^{**}$, $q_m > q_s$, $\pi_m > \pi_s$ and $\mathcal{I}_s > \mathcal{I}_m$ in which*

$$\mathcal{I}_m \equiv \frac{\Pr(\textit{both agents report} \mid \theta_1\theta_2 = 0)}{\Pr(\textit{both agents report} \mid \theta_1\theta_2 = 1)} = \frac{\Pr(\textit{agent } i \textit{ reports} \mid \theta_i = 0)}{\Pr(\textit{agent } i \textit{ reports} \mid \theta_i = 1)}, \quad \textit{for } i \in \{1,2\}.$$

The proof consists of three parts, which are in Appendix B.1 (comparison to the single-agent benchmark), Appendix B.2 (principal's incentives), and Online Appendix C (conviction probabilities). According to Theorem 1, *all* equilibria share the following properties when L is large: First, the principal abuses each agent with positive probability but never abuses both agents at the same time. Second, the principal is convicted with strictly positive probability only when two reports are filed against him. Third, compared to the single-agent benchmark, reports are less informative both at the individual and at the aggregate level. As a result, the equilibrium probability of crime is strictly higher. Importantly, strategic agents are *more* likely to file reports in the two-agent setting than in the single-agent setting. This feature distinguishes our result from the literature on public good provision, in which inefficiency is driven by the scarcity of contributions.

Our second theorem examines the informativeness of agents' reports and the equilibrium probability of crime as the punishment to benefit ratio, L, becomes arbitrarily large.

Theorem 2. *For every $\epsilon > 0$, there exists $\overline{L}_\epsilon \geq 0$ such that when $L > \overline{L}_\epsilon$, every equilibrium satisfies: $\omega_m^* < -1/\epsilon$, $\omega_m^{**} < -1/\epsilon$, $\mathcal{I}_m < 1 + \epsilon$, and $\pi_m \geq \pi^* - \epsilon$.*

The proof is in Appendix B.3. Theorem 2 suggests that regardless of the relative magnitudes of b and c, as the punishment to the convicted principal becomes arbitrarily large relative to his benefit of committing crime, agents' reports become *arbitrarily uninformative* about whether a crime has occurred or not, and the probability of crime converges to the conviction threshold π^*. This stands in sharp contrast to the single-agent benchmark, in which the agent's report becomes arbitrarily informative and the probability of crime vanishes to zero as $L \to \infty$.

Given the presence of mechanical types whose reports transmit no information, one may suspect that $\mathcal{I}_m \to 1$ is driven by the scarcity of reports filed by the strategic types. The fourth part of Theorem 1 refutes this conjecture. It says that the probability that a strategic type files an accusation is strictly higher in the two-agent case than in the single-agent benchmark: $\omega_m^* > \omega_s^*$ and $\omega_m^{**} > \omega_s^{**}$. Since the agent's report is arbitrarily informative when $L \to \infty$ in the single-agent benchmark, their reports becoming arbitrarily uninformative in the two-agent case cannot be caused by the scarcity of reports filed by the strategic types.

The above comparison between reporting thresholds generalizes to any finite number of agents (Proposition 8). It illustrates the difference between the inefficiencies arising in our paper and those that arise in games of public good provision (Chamberlin 1974). In games of public good provision, each agent contributes less when there are more agents as free-riding incentives become more severe. In our setting, each agent is more likely to file an accusation when there are more agents, and the inefficiencies are caused by the rising probability of false accusations and the lack of reporting credibility, as opposed to the agents' incentives to free-ride.

Proof Sketch: We provide some key steps towards proving these theorems, which show why these results are driven by the tension between two forces. First, a negative correlation between θ_1 and θ_2 (the agents' private information); and second, the agents' coordination motives in filing accusations. Both forces arise endogenously when L is large enough.

We begin by explaining why, when L is large enough, the principal is convicted only when he has been accused by both agents. To see this, suppose by way of contradiction that a single accusation suffices to generate a conviction with positive probability. Then the principal must be surely convicted if two accusations were leveled against him. This is because two accusations make him strictly more likely of having committed crime, compared to the case in which there is only one accusation. We show in Proposition C.1 of Online

Appendix C that the above observation implies a uniform lower bound on the marginal increase in conviction probabilities when the principal commits an extra crime. When L is large enough, this uniform lower bound implies that the principal has a strict incentive not to commit any crime, contradicting the earlier result (Lemma 3.1) that the equilibrium probability of crime is interior. Thus, we must have $q(0,0) = q(1,0) = q(0,1) = 0$ and $q(1,1) \in (0,1)$ in all equilibria.

Since the principal's benefit from committing each crime is constant, whether his choices of θ_1 and θ_2 are strategic complements or strategic substitutes depends on the sign of

$$q(1,1) + q(0,0) - q(1,0) - q(0,1), \tag{3.7}$$

which determines whether the conviction probability is convex or concave in the number of reports. We show in Lemma B.1 (Appendix B.2) that

1. If (3.7) is strictly positive, then θ_1 and θ_2 are strategic substitutes.

2. If (3.7) is strictly negative, then θ_1 and θ_2 are strategic complements.

The previous step implies that (3.7) is strictly positive under a large L. This, together with Lemma 3.1, implies that the principal is indifferent between abusing no agent and abusing only one agent. But the principal has a strict incentive not to abuse agent j once he has already abused agent i. This leads to an *endogenous negative correlation* between θ_1 and θ_2.

From the perspective of agent 1, the equilibrium conviction probabilities imply that he has more incentive to report when he believes that agent 2 is more likely to report, since it increases his chances of avoiding the retaliation cost c. Based on the principal's equilibrium strategy, if $\theta_1 = 0$, then agent 1 believes that agent 2 is abused with probability 0 and is less likely to report, and vice versa. This discourages agent 1 from reporting when he has been abused, and encourages him to report when he has not been abused. In summary, the agents have negatively correlated private information but their decisions to accuse the principal are strategic complements. These forces undermine the credibility of agents' reports and increase the probability of crime.

As $L \to \infty$, two competing effects arise. First, as in the single-agent benchmark, the probability of false accusations decreases, which improves informativeness. Second, the distance between the reporting cutoffs of abused and intact agents decreases, which is caused by the negative correlation and the agents' coordination motives. This undermines informativeness. Theorem 2 shows that the second effect dominates the first one, irrespective of the magnitude of b and c.

To better understand the interplay between these economic forces, we sketch a proof of Theorem 2 by deriving formulas for the agents' reporting cutoffs, the informativeness of reports, and the conviction probabilities.

Throughout these derivations, we take the conclusions in Theorem 1 as given. For agent $i \in \{1, 2\}$, if $\theta_i = 0$, then he prefers to report when

$$\omega \leq \omega_m^* \equiv -c\frac{1 - q_m Q_0}{q_m Q_0} = c - \frac{c}{q_m Q_0}, \tag{3.8}$$

where Q_0 is the probability that agent j ($\neq i$) accuses the principal *conditional on* $\theta_i = 0$. Similarly, if $\theta_i = 1$, then agent i prefers to report when

$$\omega \leq \omega_m^{**} \equiv -b - c\frac{1 - q_m Q_1}{q_m Q_1} = -b + c - \frac{c}{q_m Q_1}, \tag{3.9}$$

where Q_1 is the probability that agent j accuses the principal *conditional on* $\theta_i = 1$. The expressions for Q_0 and Q_1 are given by

$$Q_0 = \delta\Phi(\omega_m^{**}) + (1 - \delta)\alpha \tag{3.10}$$

and

$$Q_1 = \delta\Big(\beta\Phi(\omega_m^{**}) + (1 - \beta)\Phi(\omega_m^*)\Big) + (1 - \delta)\alpha, \tag{3.11}$$

where

$$\beta \equiv \frac{1 - \pi_m}{1 - \pi_m/2} \tag{3.12}$$

is the probability that $\theta_j = 1$ conditional on $\theta_i = 1$.

The comparisons between Q_0 and Q_1 and between ω_m^* and ω_m^{**} unveil the key difference between the two-agent scenario and the single-agent benchmark. Instead of having a constant distance b, the distance between the reporting cutoffs is given by:

$$\omega_m^* - \omega_m^{**} = b - \frac{c}{q_m} \cdot \frac{-1 + Q_1/Q_0}{Q_1}. \tag{3.13}$$

This leads to the following lemma:

Lemma 3.3. $\omega_m^* - \omega_m^{**} \in (0, b)$.

Proof. From (3.10) and (3.11), $\omega_m^* - \omega_m^{**} > 0$ is equivalent to $Q_1 > Q_0$. To see this, suppose by way of contradiction that $Q_1 \leq Q_0$. Equation (3.13) then implies that $\omega_m^* \geq \omega_m^{**} + b > \omega_m^{**}$. The comparison between (3.8) and (3.9) then yields $Q_1 > Q_0$, the desired contradiction. Since $Q_1 > Q_0$, the term $\frac{-1+Q_1/Q_0}{Q_1}$ is strictly positive. Therefore, $\omega_m^* - \omega_m^{**} < b$. □

Next, we explore how the decrease in $\omega_m^* - \omega_m^{**}$ affects the informativeness of reports and the equilibrium probability of crime. First, we provide a formula for the informativeness ratio $\mathcal{I}_m$, which has been defined in

the fourth statement of Theorem 1:

$$\mathcal{I}_m \equiv \frac{\big(\delta\Phi(\omega_m^*) + (1-\delta)\alpha\big)\big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\big)}{\big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\big)^2} = \frac{\delta\Phi(\omega_m^*) + (1-\delta)\alpha}{\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha}. \tag{3.14}$$

Since $q_m \in (0,1)$, the evaluator believes that $\theta_1\theta_2 = 0$ with probability π^* after observing two reports. Therefore, the equilibrium probability of crime π_m satisfies:

$$\frac{\pi_m}{1-\pi_m} = \frac{l^*}{\mathcal{I}_m}, \quad \text{in which} \quad l^* \equiv \frac{\pi^*}{1-\pi^*}. \tag{3.15}$$

Plugging (3.15) into (3.12), we have the following expressions for β and $1-\beta$:

$$\beta = \frac{2\mathcal{I}_m}{l^* + 2\mathcal{I}_m} \text{ and } 1-\beta = \frac{l^*}{l^* + 2\mathcal{I}_m}. \tag{3.16}$$

Plugging (3.16) into (3.10) and (3.11), we obtain the following expression for the ratio between Q_1 and Q_0:

$$\frac{Q_1}{Q_0} = \beta + (1-\beta)\mathcal{I}_m = \frac{(l^*+2)\mathcal{I}_m}{l^* + 2\mathcal{I}_m}. \tag{3.17}$$

Plugging (3.8) and (3.9) into (3.17), we obtain

$$\frac{|\omega_m^* - c|}{|\omega_m^{**} - c + b|} = \frac{-c/q_m Q_0}{-c/q_m Q_1} = \frac{Q_1}{Q_0} = \frac{(l^*+2)\mathcal{I}_m}{l^* + 2\mathcal{I}_m}. \tag{3.18}$$

This leads to the following lemma:

Lemma 3.4. *If $\omega_m^* \to -\infty$, then $\mathcal{I}_m \to 1$ and $\pi_m \to \pi^*$.*

Proof. Since $\omega_m^* - \omega_m^{**} \in (0, b)$, the difference between $|\omega_m^* - c|$ and $|\omega_m^{**} - c + b|$ is at most b. The LHS of (3.18) converges to 1 as $\omega_m^* \to -\infty$. Since the RHS of (3.18) is strictly increasing in $\mathcal{I}_m$, we know that the limiting value of $\mathcal{I}_m$ is 1. According to (3.15), the limiting value of π_m is π^*. □

To complete this heuristic proof of Theorem 2, we now argue that ω_m^* and ω_m^{**} converge to $-\infty$ as $L \to \infty$. This is driven by the principal's incentive constraint: the principal must be indifferent between committing one crime and committing no crime:

$$\frac{1}{\delta L} = q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big). \tag{3.19}$$

If ω_m^{**} converges to some finite number ω^{**}, then either $q_m \to 0$ or $|\omega_m^* - \omega_m^{**}| \to 0$. When $q_m \to 0$, (3.8) and (3.9) suggest that both ω_m^* and ω_m^{**} converge to $-\infty$. This contradicts the hypothesis that ω_m^{**} converges

to some finite number. When $|\omega_m^* - \omega_m^{**}| \to 0$ and ω_m^{**} converges to some finite number, then $\mathcal{I}_m \to 1$ and $Q_1/Q_0 \to 1$. Expression (3.13) then implies that the limiting value of $\omega_m^* - \omega_m^{**}$ is b. This contradicts the hypothesis that $\omega_m^* - \omega_m^{**}$ converges to 0.

4 Restoring the Credibility of Reports

Focusing on the case with two agents, we provide three solutions to restore the credibility of reports and to reduce the probability of crime. First, we show that the probability of crime is lower under an intermediate level of punishment. Second, we construct transfers to the agents to restore their reporting credibility. Third, we consider an alternative conviction rule, which concerns how to apply the threshold conviction probability.

4.1 Lower Punishment

We show that under some intermediate levels of punishment, the agents' reports are more credible and the equilibrium probability of crime is lower, compared to the case in which the punishment is arbitrarily high.

Proposition 4. *For every $c > 0$, there exists an interval $[\underline{L}(c), \overline{L}(c)]$ such that an equilibrium exists when $L \in [\underline{L}(c), \overline{L}(c)]$. In every equilibrium, the principal either abuses two agents or abuses no agent, and the conviction probabilities satisfy:*

$$q(1,1) + q(0,0) - q(1,0) - q(0,1) < 0.$$

For every $\epsilon > 0$, there exists $c_\epsilon > 0$ such that when $c > c_\epsilon$ and $L \in [\underline{L}(c), \overline{L}(c)]$, there exists an equilibrium in which the probability of crime is less than ϵ.

The proof is in Online Appendix D. Proposition 4 shows that there exists an intermediate range of L, under which the conviction probabilities are *concave* in the number of reports in all equilibria. As a result, the principal's decisions are strategic complements. In equilibrium, he either abuses no agent or abuses both agents. Moreover, abusing a single agent is strictly suboptimal.

The principal's equilibrium strategy induces a *positive correlation* between θ_1 and θ_2. In contrast to when L is large, each agent's coordination motive now encourages him to report when he has witnessed a crime and vice versa. This increases the distance between his two reporting thresholds, making it strictly larger than b. As a result, the informativeness of reports increases and the probability of crime decreases. As each agent's loss from miscoordination becomes arbitrarily large, each individual report becomes arbitrarily informative and the equilibrium probability of crime converges to zero.

Proposition 4 implies that in order to minimize the probability of crime, the optimal magnitude of punishment is *interior* when there are multiple potential witnesses who are vulnerable to retaliation. This finding differs from Becker's (1968) seminal analysis of criminal justice and law enforcement, which suggests that increasing the magnitude of punishment helps reduce crime. From this perspective, our finding provides a novel rationale for being lenient to the convicted. Our logic applies to settings in which smoking-gun evidence is scarce and the potential victims' claims are hard to verify. Importantly, reducing L comes at the cost of increasing the number of crimes conditional on the principal being guilty. This reveals a trade-off between reducing the probability of crime and reducing the number of crimes.

4.2 Monetary Transfers

In this section, we maintain the assumption that L is large and explore the use of monetary transfers to mitigate the coordination problem across agents. The transfers are contingent on the vector of reports: we let $t_i(\boldsymbol{a}) \in \mathbb{R}$ denote the transfer to agent i under reporting profile $\boldsymbol{a}$. We begin by constructing a transfer scheme that eliminates coordination inefficiencies. Let

$$t_1^*(a_1, a_2) = \begin{cases} c & \text{if } (a_1, a_2) = (1, 0) \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad t_2^*(a_1, a_2) = \begin{cases} c & \text{if } (a_1, a_2) = (0, 1) \\ 0 & \text{otherwise} \end{cases}$$

Under transfer scheme (t_1^*, t_2^*), the next proposition shows that agents' reports become arbitrarily informative and the probability of crime vanishes to zero as $L \to \infty$.

Proposition 5. *For every $\epsilon > 0$, there exists $\overline{L}_\epsilon > 0$ such that for all $L > \overline{L}_\epsilon$ and under transfer scheme (t_1^*, t_2^*), the informativeness ratio of each agent's report exceeds $1/\epsilon$ and the probability of crime is less than ϵ in all equilibria.*

The proof is in Online Appendix E.1. Proposition 5 implies that when it is hard to scale down punishment (e.g., conviction has other negative consequences on one's career that is beyond the control of the judge), one can restore the informativeness of reports by compensating lone accusers. The amount to be transferred exactly offsets an agent's loss from retaliation, which eliminates the agents' incentives to coordinate. As a result, the distance between each agent's reporting cutoffs is b. However, the equilibrium outcome does not coincide with that of the single-agent benchmark: the principal's incentives are different, and the cutoffs under the two-agent setting with transfer scheme (t_1^*, t_2^*) are strictly higher than those arising in the single-agent benchmark. For any given L, the informativeness of reports is strictly lower and the probability of crime is strictly higher than in the single-agent setting.

The above transfer scheme presents two weaknesses. First, the designer needs to incur a budget deficit with positive probability. This deficit is increasing in the agents' losses from retaliation. Second, this scheme encourages collusion between the principal and the agents. For example, the principal and agents can agree that only agent 1 reports and the principal does not retaliate. After agent 1 obtains the transfer c, he will share it with agent 2 and the principal.

Motivated by these concerns—as well as by appropriately positioning our work relative to Crémer and McLean (1985, 1988)—we explore the possibility of restoring reporting informativeness via budget- balanced transfer schemes, namely, transfers such that $\sum_{i=1}^{2} t_i(\boldsymbol{a}) = 0$ for every $\boldsymbol{a} \in \{0,1\}^2$. We allow t to be asymmetric across agents and also allow for asymmetric equilibria under these transfer schemes. In Online Appendix E.2, we show that the informativeness of agents' reports is uniformly bounded from above when L is large. This finding reveals a tension between improving informativeness, preventing any budget deficit, and deterring collusion.

Intuitively, this impossibility result is driven by the tension between two objectives: (1) mitigating the adverse effect of agents' coordination motives when the punishment L is large, and (2) deterring false accusations. The former objective is achieved by increasing $\Delta_1 \equiv t_1(1,0) - t_1(0,0)$ and $\Delta_2 \equiv t_2(0,1) - t_2(0,0)$, while the latter objective is achieved by decreasing ω_1^{**} and ω_2^{**}, i.e., by reducing agents' incentives to file false accusations. The budget-balance requirement forbids the principal from achieving both goals simultaneously, which causes bounded informativeness and a non-vanishing probability of crime.

4.3 Alternative Conviction Rules

We consider an alternative conviction rule based on a different kind of the conviction threshold. Let $\pi_i \equiv \mathbb{E}[1-\theta_i|\boldsymbol{a}]$ denote the posterior probability that the principal is guilty of abusing agent i conditional on agents' reports. Suppose, focusing for simplicity on two agents, that the evaluator chooses the conviction decision $s \in \{0,1\}$ as follows:[17]

$$s \begin{cases} = 0 & \text{if} \quad \max_{i\in\{1,2\}} \pi_i > \pi^* \\ \in \{0,1\} & \text{if} \quad \max_{i\in\{1,2\}} \pi_i = \pi^* \\ = 1 & \text{if} \quad \max_{i\in\{1,2\}} \pi_i < \pi^*, \end{cases} \tag{4.1}$$

Recall that in the single-agent benchmark, the reporting thresholds are ω_s^* and ω_s^{**}, the conviction probability in case an agent reports is q_s and the equilibrium probability of crime is π_s, all of which depend on L. We have the following result, which can be generalized to any finite number of agents.

[17] A more straightforward solution is to punish the principal by mL when $\#\{i|\pi_i \geq \pi^*\} = m$. However, this solution is at odds with the situations that motivate our analysis, in which punishment is by nature discrete and not easily scalable with respect to the number of crimes. This is the case when conviction entails the loss of one's job, power, or influence.

Proposition 6. *Under conviction rule (4.1), for L large enough, there exists equilibrium in which:*

1. *The principal abuses each agent with probability π_s and the probability with which he abuses agent i is independent of whether he abuses agent j.*

2. *The principal is convicted with probability mq_s where $m \in \{0, 1, 2\}$ is the number of accusations filed against him.*

3. *Agent i accuses the principal if and only if $\omega_i \leq \omega_s^*$ and $\theta_i = 0$, or $\omega_i \leq \omega_s^{**}$ and $\theta_i = 1$.*

4. *The equilibrium probability of crime is $1 - (1 - \pi_s)^2$.*

In the limit where $\delta \to 1$ and $L \to \infty$, the equilibrium probability of crime goes to 0.

The takeaway message is that using $\max_{i \in \{1,2\}} \pi_i$, instead of the principal's overall probability of guilt, as the basis for conviction removes the negative correlation in the agents' private information and restores the credibility of their accusations. This insight is not driven by the assumption that the principal can be convicted with any arbitrary probability at the conviction threshold π^*. In general, under any increasing function mapping $\max_{i \in \{1,2\}} \pi_i$ to a level of punishment, one can show that agents' observations of crime cannot be negatively corrected in any equilibrium that respects presumption of innocence (Axioms 1) and a monotonicity axiom under which the expected punishment to the principal is nondecreasing when a larger set of agents report.

Our results provide a rationale for treating the probability of each crime separately, rather than focus on the principal's overall probability of guilt. Nevertheless, such a conviction rule is hard to implement in practice if the evaluator lacks commitment. To illustrate, consider the comparison between the following two defendants:

-	Prob of abuse victim 1	Prob of abuse victim 2	Prob of abusing at least one victim
Defendant 1	49 %	49 %	98 %
Defendant 2	50 %	1 %	51 %

Under conviction rule (4.1), defendant 1 is acquitted for sure if defendant 2 is acquitted, while common sense suggests that most judges are more likely to convict defendant 1. In context of organizations, for instance, a firm faces more social pressure to fire a manager whose probabilities of committing abuse correspond to defendant 1, compared to a manager whose probabilities of committing abuse match defendant 2.

5 Extensions

Section 5.1 extends our theorems to three or more agents. Section 5.2 introduces heterogeneity across principals: the principal's cost to benefit ratio from committing crime is stochastic and is his private information. Section

5.3 explores several variants of the baseline model under which our insights remain robust, which include the availability of ex post evidence, the agents and the evaluator facing uncertainty about the number of potential victims, the principal and the agents having alternative payoff functions, and the mechanical types' reports being informative about the true state.

5.1 Arbitrary Number of Agents

When there are two or more agents, we say that an equilibrium is *unanimous* if $q(\boldsymbol{a}) = 0$ unless $\boldsymbol{a} = (1, 1, ..., 1)$, namely, no conviction unless agents unanimously report. We have the following result:

Proposition 7. *For every $n \geq 2$, there exists $\overline{L}_n \in \mathbb{R}_+$ such that when $L > \overline{L}_n$,*

1. *There exists a symmetric unanimous equilibrium that satisfies Axiom 1.*

2. *Every symmetric equilibrium that satisfies Axiom 1 is a unanimous equilibrium. In this equilibrium, the principal abuses at most one agent.*

For every $\epsilon > 0$, there exists $\overline{L}_{n,\epsilon} \geq \overline{L}_n$ such that when $L > \overline{L}_{n,\epsilon}$, in every symmetric equilibrium that satisfies Axiom 1, each agent's reporting cutoffs are less than $-1/\epsilon$, the aggregate informativeness of reports is below $1 + \epsilon$, and the equilibrium probability of crime is more than $\pi^* - \epsilon$.

The proof is in Online Appendix F. When the principal faces an arbitrarily large punishment, one can show that the principal is convicted only when agents unanimously report. When conviction probabilities possess this property, the principal's marginal cost of abusing an additional agent is increasing in the number of agents abused. In equilibrium, either no agent is abused or only one agent is abused. The principal's equilibrium strategy induces a negative correlation in the agents' private information. This, together with the agents' incentive to coordinate their report, discourages them from accusing the principal when they have been abused and vice versa.

Previous sections have established that moving from a single agent to two agents could significantly reduce informativeness and increase the probability of crime. We now extend this result by studying the comparative statics of equilibria with respect to the number of agents, focusing on symmetric unanimous equilibria. For every $n \in \mathbb{N}$, let ω_n^* and ω_n^{**} be an agent's reporting cutoffs when he has been abused and when he has not been abused, respectively. Also let π_n denote the equilibrium probability of crime and

$$\mathcal{I}_n \equiv \frac{\Pr(n \text{ agents report}|\ \Pi_{i=1}^n \theta_i = 0)}{\Pr(n \text{ agents report}|\ \Pi_{i=1}^n \theta_i = 1)}$$

denote the aggregate informativeness of the agents' reports. Since at most one agent is abused in any unanimous equilibrium,

$$\mathcal{I}_n = \frac{\delta\Phi(\omega_n^*) + (1-\delta)\alpha}{\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha}.$$

Let $\overline{L}_n \in \mathbb{R}_+$ be such that for every $L > \overline{L}_n$, there exists a symmetric unanimous equilibrium in an n-agent setting.

Proposition 8. *For every $k > n$ and $L > \max\{\overline{L}_n, \overline{L}_k\}$, the following inequalities hold for any symmetric unanimous equilibria with k and n agents: $\omega_k^* > \omega_n^*$, $\omega_k^{**} > \omega_n^{**}$, $\omega_k^* - \omega_k^{**} < \omega_n^* - \omega_n^{**}$, $\pi_k > \pi_n$ and $\mathcal{I}_n > \mathcal{I}_k$.*

The proof is in Appendix B.4. Proposition 8 shows that as the number of agents increases, the distance between the reporting cutoffs decreases and the aggregate informativeness of the agents' reports also decreases. In equilibrium, this leads to a higher probability of crime. Moreover, each agent is more likely to report in an environment with more agents. The driving force behind such comparative statics is still the interaction between the coordination motives among agents and the negative correlation between their private information.

5.2 Uncertain Principal Preferences

We now examine the robustness of our insights when there is uncertainty about the principal's preferences. We consider the presence of (1) *virtuous* principals, whose cost from the commission of crime is high relative to the utility that they gain from it (e.g., these individuals have more to lose due to major career concerns), or whose benefit from committing crimes is zero or negative; and (2) *vicious* principals, who experience a high utility from the commission of crime or face a lower punishment in case they are found guilty relative to opportunistic type principals.

The principal's preference type is privately observed. Still normalizing the benefit of crime to 1, we model heterogeneity by treating the punishment cost $\widetilde{L}$ to be stochastic, instead of the constant L used in the baseline model. For simplicity, we focus on environments in which there are two agents and the random variable $\widetilde{L}$ takes two possible values, L_l and L_h, with $L_l < L_h$. We consider two cases separately. These two cases together establish the robustness of our results to the presence of virtuous and vicious types, as long as both of these types occur with low enough probability.

Virtuous principals: Suppose that L_l is larger than the lower bound on L required by Theorem 1 and the probability with which $L = L_h$, denoted by π_h, is strictly less than $1 - \pi^*$. All equilibria in this environment take a similar form as those in Theorem 1, which are characterized in Proposition 9:

Proposition 9. *There exists $\overline{L} \in \mathbb{R}_+$ such that when $L_l > \overline{L}$ and $\pi_h < 1 - \pi^*$, every equilibrium is characterized via $\{\omega^*, \omega^{**}, q, \pi\}$ with $\pi < \pi^*$ such that*

1. *Agent i accuses the principal when $\{\omega_i \leq \omega^*$ and $\theta_i = 0\}$ or $\{\omega_i \leq \omega^{**}$ and $\theta_i = 1\}$.*

2. *Type L_l principal chooses $(\theta_1, \theta_2) = (1, 1)$ with probability $\frac{1-\pi}{1-\pi_h}$, $(\theta_1, \theta_2) = (1, 0)$ with probability $\frac{\pi}{2(1-\pi_h)}$, and $(\theta_1, \theta_2) = (0, 1)$ with probability $\frac{\pi}{2(1-\pi_h)}$.*

3. *The conviction probabilities satisfy $q(0, 0) = q(0, 1) = q(1, 0) = 0$ and $q(1, 1) = q$.*

Moreover, for every $\epsilon > 0$, there exists $\overline{L}_\epsilon > \overline{L}$ such that when $L_l > \overline{L}_\epsilon$, the informativeness ratio of each individual report is less than $1 - \epsilon$ and the equilibrium probability of crime is more than $\pi^* - \epsilon$.

The proof is only a minor departure from those of Theorems 1 and 2 and is omitted. Proposition 9 shows that the presence of a virtuous type increases the opportunistic type's probability of committing crimes. The intuition is simple: If the fraction of virtuous principal increases, this lowers the judge's assigned probability that the principal is guilty, other things equal. This change decreases the expected punishment to the principal, which provides cover for the opportunistic type principal and encourages him to commit crimes. This adjustment yields a new equilibrium in which a larger fraction of opportunistic types commit crimes and the fraction of guilty principals (unconditional on their types) in the population matches the previous equilibrium level.

Vicious principals: Suppose now that L_h is larger than the lower bound on L required by Theorem 1 while L_l is strictly less than 1.[18] We assume that $\widetilde{L} = L_l$ with a small enough probability ϵ. In equilibrium, a principal of type L_l behaves as a *serial abuser*, with a high propensity to commit crimes. Type L_h is interpreted as the *opportunistic types* who enjoy moderate benefit from committing crimes and strategically decides whether or not to commit crimes as well as how many crimes to commit. The following result demonstrates the robustness of the endogenous negative correlation between θ_1 and θ_2, which is the key driving force behind the coordination inefficiencies identified by our results:

Proposition 10. *For each $\epsilon > 0$ small enough, there exists $\overline{R} > 1$ such that for every $R \in (1, \overline{R})$, there exists L_h as well as a symmetric equilibrium $\{\omega^*, \omega^{**}, \pi, q\}$ under (ϵ, L_h, L_l), such that:*

1. *The conviction probabilities are $q(0, 0) = q(1, 0) = q(0, 1) = 0$ and $q(1, 1) = q$.*

2. *Type L_l principal abuses both agents for sure. Type L_h assaults agent i with probability $\frac{\pi-\epsilon}{2(1-\epsilon)}$, for every $i \in \{1, 2\}$, and assaults no agent with probability $(\pi - \epsilon)/(1 - \epsilon)$.*

3. *The agents' reporting cutoffs satisfy $\frac{\delta\Phi(\omega^*)+(1-\delta)}{\delta\Phi(\omega^{**})+(1-\delta)} = R$.*

[18] Assuming that $L_l < 1$ facilitates exposition. For our result to hold, we require that L_l be sufficiently small relative to L_h so that a principal with type L_l has an incentive to commit two crimes in equilibrium.

4. *The equilibrium probability of crime satisfies* $\pi = \frac{l^*}{l^*+\mathcal{I}}$. *in which* $\mathcal{I}$ *is a measure of report informativeness, defined as:*

$$\mathcal{I} \equiv \frac{\epsilon}{\pi}R^2 + \Big(1 - \frac{\epsilon}{\pi}\Big)R.$$

The proof is in Online Appendix G. According to Proposition 10, when the probability of vicious type is small enough, there exist a sequence of punishments to the opportunistic type as well as a sequence of equilibria under these parameters, such that the agents' reports are arbitrarily uninformative at the individual level (measured by R), and at the aggregate level (measured by $\mathcal{I}$). The latter leads to a high probability of crime, and in particular, $\frac{l^*}{l^*+\mathcal{I}}$ converges to π^* as $\mathcal{I}$ becomes close to 1.

Intuitively, introducing a small probability of serial abusers cannot overturn the negative correlation between the agents' private information, which is driven by the opportunistic type principal's strategic restraint when L_h is large. Just as in the baseline model, an agent who has been abused holds a pessimistic belief about the other agent's reporting probability. Each agent's coordination motive in filing accusations hurts his reporting credibility and in equilibrium, this increases the probability of crime committed by the opportunistic type.

While Theorems 1 and 2 describe properties of all equilibria (subject to the symmetry and presumption of innocence refinements), Proposition 10 demonstrates only the existence of *some* equilibrium in which θ_1 and θ_2 being negatively correlated, the informativeness of reports being close to 1 and the probability of crime being close to π^*. The presence of vicious types can lead to a larger variety of self-fulfilling beliefs with different qualitative features, even when those types occur with arbitrarily small probability.

To understand why, let us start from an agent's equilibrium strategy in an economy without vicious types, that is, $\epsilon = 0$. This is characterized by two cutoffs (ω^*, ω^{**}). When ϵ is strictly positive, it undermines the negative correlation between θ_1 and θ_2, which encourages agent i to report when $\theta_i = 0$ and discourages him from reporting when $\theta_i = 1$. The increase in the distance between ω^* and ω^{**} then increase the informativeness of reports and in equilibrium, it decreases the net probability of crime. Since the probability of the principal being a serial assaulter is fixed to be ϵ, a decrease in the total probability of crime increases

$$\Pr(\theta_1 = \theta_2 = 0 | \theta_1\theta_2 = 0). \tag{5.1}$$

This further weakens the negative correlation between θ_1 and θ_2, which increases the informativeness of report and decreases the probability of crime. One can iterate the above argument until it reaches a new fixed point. The equilibrium probability of crime could be close to ϵ, or close to π^*, or somewhere in between, depending on the starting point.

5.3 Other Extensions

Decreasing marginal benefits from crime: Our result remains robust when the principal faces decreasing marginal returns from committing multiple crimes (Becker 1968) or receives a punishment larger than L when he is believed to have committed multiple crimes. These changes motivate the principal to commit fewer crimes and induce, as in the baseline model, negative correlation in the agents' private information. As in the baseline model, the agents' coordination motives undermine the informativeness of their reports and increase the probability of crime. In Online Appendix H.3, we study an extension of the baseline model that formalizes these arguments. In particular, the principal is convicted of a minor crime and receives punishment L if the probability with which he is guilty of at least one crime exceeds π^*, and is convicted of a felony and receives punishment $L'(> L)$ if the probability with which he is guilty of two crimes exceeds π^{**}. When L is large (e.g., the principal can lose his lucrative position when convicted with a minor offense), the principal commits at most one crime in every symmetric equilibrium that satisfies presumption of innocence (Axiom 1).

Agent's Reporting Cost: Our results extend when each agent's suffers a lower retaliation cost when there are more reports filed against the principal, as long as the retaliation cost is strictly positive whenever the principal is acquitted. This variation strengthens the coordination motives among agents without affecting the negative correlation between their private information.

Agent's Interdependent Preferences: Agent i's payoff function could directly depend on other agents types θ_j. This will be the case, for instance, if agents have social preferences and directly care about crimes committed against (or observed by) other agents.

We capture this possibility by modifying agent i's payoff as follows:

$$\Big\{ \underbrace{\omega_i + b\Big(\gamma\theta_i + (1-\gamma)\Pi_{j=1}^{n}\theta_j\Big) - ca_i}_{\text{payoff when the principal is acquitted}} \Big\} s. \tag{5.2}$$

Agent i's payoff when the principal is convicted ($s = 0$) is still normalized to 0, but his payoff when the principal is acquitted ($s = 1$) depends not only on whether i has been abused, but also on whether the principal has committed crimes against other agents. The parameter $\gamma \in [0, 1]$ is the weight attached to θ_i (whether the principal committed a crime against him or not) relative to the weight attached on $\Pi_{j=1}^{n}\theta_j$ (whether the principal is guilty at all).

As in the baseline model, when L is large, agents' reports become arbitrarily uninformative and the equilibrium probability of crime approaches π^* (Online Appendix H.1). In fact, social preferences can reduce informativeness,

because agent i's report becomes more responsive to his belief about θ_j instead of to his observation of θ_i. Since θ_i and θ_j are negatively correlated in equilibrium, this dampens the responsiveness of agent i's reporting strategy to θ_i, making it less informative.

Ex Post Evidence & Punishing False Accusations: We consider the possibility that evidence may arise ex post, which exposes false accusations. For example, suppose that when an innocent principal is convicted, hard evidence arrives with probability p^* that reveals his innocence, causing every false accuser to be penalized by some constant $\ell \geq 0$. Our analysis is essentially unchanged, because such punishments are equivalent to an increase in the added benefit b from reporting after witnessing a crime. This extension is formally considered in Online Appendix H.1.

Uncertainty about the Number of Potential Victims: In applications such as workplace bullying, physical assaults and discrimination, the number of potential victims is usually not observed by the judge and the victims. Motivated by this concern, we consider an extension of our model, in which nature randomly a subset $\widetilde{N}$ of $\{1, 2, ..., n\}$, interpreted as the set of agents the principal has opportunities to abuse. We assume that only agents in $\widetilde{N}$ can be assaulted and file reports, the latter assumption being interpreted as follows: if an agent outside of $\widetilde{N}$ files an accusation, this accusation can be easily refuted by the defendant (e.g., using an alibi). Only the principal observes $\widetilde{N}$. Agent i privately observes whether $i \in \widetilde{N}$ or not in addition to his private information in the baseline model.

We informally argue that the economic mechanisms behind our results are stronger when the judge and the agents face this extra layer of uncertainty. Since the judge does not observe the size of $\widetilde{N}$, whether the principal is convicted or not depends only on the number of reports, but not on the number of potential victims. Since the principal is convicted with weakly higher probability when there are more reports, he has stronger incentives to commit assaults when fewer agents can report (that is, $|\widetilde{N}|$ is smaller). Conditional on an agent being assaulted, the agent infers that $|\widetilde{N}|$ is more likely to be small and, hence, the expected number of reports filed by other agents is also likely to be smaller, other things equal. This dampens the abused agents' incentive to report and lowers the agent's credibility in equilibrium, by the same logic as in the baseline model.

Mechanical Types' Strategies: Suppose, in contrast to the baseline model, the mechanical type agent accuses the principal with probability $\overline{\alpha}$ when he has observed a crime, and accuses the principal with probability $\underline{\alpha}$ otherwise, with $1 > \overline{\alpha} \geq \underline{\alpha} > 0$. This formulation of mechanical types incorporates strategic types who are immune to the principal's retaliation. This is because without retaliation, a strategic agent maximizes $(\omega_i + b\theta_i)s$. In equilibrium, his reporting cutoffs are 0 and $-b$, depending on whether he has been abused or

not. The conditional probabilities with which he reports are $\overline{\alpha} = \Phi(0)$ and $\underline{\alpha} = \Phi(-b)$, respectively. Therefore, he behaves equivalently to a mechanical type that plays an informative cutoff-strategy.

We extend our results to this environment in Online Appendix H.2. We show that no matter how informative the mechanical types' reports are, it will be overturned by the strategic type agent's coordination motives when δ is close to 1 and L is large. As a result, the agents' reports become arbitrarily uninformative as $L \to \infty$.

6 Conclusion

Beyond the different modeling variations discussed in section 5, our results may be viewed from a broader perspective. We address some more fundamental issues concerning the implicit assumptions of our analysis.

Equilibrium Analysis *vs.* Nonequilibrium Adjustments: Our results are derived from an *equilibrium analysis*. This methodology, which is ubiquitous in economics, is best suited when all actors understand both the rules of the game, including the payoff consequences of their actions, and other actors' strategic motives.[19] When social rules change, as in the case of a sudden crackdown on specific crimes, the introduction of new laws and regulations, drastic shifts in social norms, or the emergence of new social media that change the social consequences of one's actions, equilibrium analysis may be viewed as a potential harbinger of issues that will emerge as economic and social actors learn to interact under these new rules.

Trade-offs in Designing Conviction Rules: Our framework and analysis suggest several trade-offs concerning the design of conviction rules for nonverifiable crimes. First, we characterize a stark tradeoff between deterrence and fairness. In particular, although the probability of crime decreases as the conviction threshold π^* decreases, this benefit comes at the cost of increasing the fraction of innocent defendants who are convicted. This is because when there are multiple potential victims and the conviction punishment being large, the posterior probability that a convicted principal is guilty is approximately π^*, and the fraction of innocent people among those that are convicted is approximately $1 - \pi^*$. Although it is intuitive, this tradeoff had largely been ignored in law and economics until Miceli (1991).[20]

Importantly, however, we have also shown that this tradeoff is not as obvious as it seems: Proposition 4 shows that using a more lenient sentence in case of conviction can paradoxically be an effective way to deter

[19] Foundations of Nash equilibrium based on players learning one another's strategies have a long history in economics. See for example Fudenberg and Levine (1995) for an in-depth discussion.

[20] One exception is Harris (1970), who introduces a reduced-form cost of judicial error. Becker (1968) and Landes (1970) ignore wrongful convictions. In Kaplow (2011), punishments are expressed in terms of fines, i.e., zero sum transfers that do not affect the social surplus. Kaplow considers the "chilling" effect of punishment on behavior. Siegel and Strulovici (2018) provide a framework that includes both deterrence and fairness considerations. Klement and Neeman (2005) study the design and cost of settlement procedures for civil cases, taking into account their effect on deterrence.

crimes, without increasing the probability of convicting the innocent. Thus, the tradeoff between deterrence and fairness depends on which instrument is considered: the conviction threshold or the sentence.

Second, suppose that the judge can *commit* to convicting the principal if at least one report is filed. Such a commitment eliminates the agents' coordination motive. When the punishment L is large, the principal has a strict incentive not to commit crime. In order to fulfill his promise, however, the judge must convict the principal after receiving any report, even though the principal is surely innocent. This leads to an undesirable outcome since all convicted individuals are innocent and the probability of convicting the innocent is significant.[21]

Shielding Accusers from Stigma through Secret Accusations: To address the potential pressure that is sometimes experienced by lone accusers, institutions have been developed under which reports are submitted to a third party and are only released when enough of them have been filed.[22]

It must be noted that, taken at face value, such institutions also protect *wrongful* accusers from stigma. Indeed, an agent holding a grudge against the principal has an opportunity to secretly file a report against the principal in the hope that other agents, rightfully or not, will also accuse the principal. While these institutions are clearly well intentioned and worth considering, it is also important to evaluate their long-term reliability.

In some cases, accusations may be leaked to the principal due to corruption, imperfect institutions, and other reasons. This risk is especially high if the principal is powerful and well connected. Our results apply when such leakages occur with strictly positive probability: the acquitted principal can retaliate against the reporting agents once the information is leaked. This is because an agent's *expected* loss from retaliation remains strictly positive and our theorems do not require conditions on the magnitude of c.

Sequential Reporting: The economic forces behind our results are also present in dynamic versions of our model, in which reports may be filed sequentially. First, the negative correlation between the agents' private information (θ_i) continues to arise endogenously whenever a strategic principal is concerned about having too many reports made against him. Second, an individual agent has an incentive to coordinate with other agents whenever he is unsure about whether his report is pivotal or not. In a dynamic setting, this incentive can materialize after a *cold start* (i.e., where very few people have reported before and no agent wants to be the first accuser). It can also occur when an agent has observed many reports and is unsure of the number of reports needed to convict the principal (for example, if he faces uncertainty about the conviction standard π^* used by the judge). The inefficiencies and lack of credibility caused by the agents' coordination motives thus still arise

[21] This paradox induced by commitment commonly arises in plea bargaining models, in which agents who reject pleas and are found guilty at trial are known to be innocent. See Grossman and Katz (1983), Reinganum (1988), and Siegel and Strulovici (2018).

[22] In particular, the nonprofit organization Callisto has a "match" feature, whereby a report is made official only if at least two victims name the same perpetrator. See www.projectcallisto.org.

in a dynamic environment.[23]

[23] See Lee and Suen (2018) for a model of strategic accusation in which the timing of accusation plays a major role.

A Equilibrium Refinements

We introduce two axioms: *monotonicity* and *properness*. These together with Axiom 1 ensure that all sequential equilibria are *symmetric* and moreover, satisfy the properties in Theorem 1.

Axiom 2 (Monotonicity)**.** *For every $\boldsymbol{a}, \boldsymbol{a'} \in \{0,1\}^n$ with $\boldsymbol{a} \succ \boldsymbol{a'}$, we have $q(\boldsymbol{a}) \geq q(\boldsymbol{a'})$.*

Axiom 2 endows the agents' reports with meanings. In particular, filing a report is a move against the principal. As a result, an agent is more likely to report when he has been abused (i.e., θ_i is low), and when he hates the principal (i.e., ω_i is small). It resonates the economic interpretation of c, which is the agent's loss from the principal's retaliation. This is because when the principal can optimally commit to retaliation plans (privately) against each agent before the game starts as in Chassang and Padró i Miquel (2018), he would commit to retaliate to the maximum against messages that increase the evaluator's belief about $\Pi_{i=1}^n \theta_i = 0$ and would not retaliate against other messages.

In every equilibrium that satisfies Axioms 1 and 2, each agent's equilibrium strategy is characterized by two cutoffs: ω_i^* and ω_i^{**}, such that agent i reports when $\omega_i \leq \omega_i^*$ and $\theta_i = 0$, or when $\omega_i \leq \omega_i^{**}$ and $\theta_i = 1$. Therefore, an equilibrium is characterized by $\left\{(\omega_i^*, \omega_i^{**}, \rho_i)_{i=1}^n, \boldsymbol{\pi}, q\right\}$, in which:

1. $\rho_i : \{0,1\} \to \Delta\left(\{0,1\}^{n-1}\right)$ is agent i's belief about θ_{-i} after observing θ_i;
2. $\boldsymbol{\pi} \in \Delta\left(\{0,1\}^n\right)$ is the principal's strategy;
3. $q : \{0,1\}^n \to [0,1]$ is the mapping from reporting profiles to conviction probabilities.[24]

Our second axiom introduces a regularity condition on agents' beliefs at off-path information sets.

Axiom 3 (Properness)**.** *For every $i \in \{1, ..., n\}$, $\widehat{\theta}_i \in \{0,1\}$, and $\theta'_{-i}, \theta''_{-i} \in \{0,1\}^{n-1}$, if the principal's expected payoff from $(\widehat{\theta}_i, \theta'_{-i})$ is strictly larger than his expected payoff from $(\widehat{\theta}_i, \theta''_{-i})$, then $\rho_i(\widehat{\theta}_i)$ attaches zero probability to θ''_{-i}.*

Axiom 3 requires that at every information set, on and off path, each agent believes that the principal is significantly less likely to make mistakes that are strictly more costly. That is to say, within the subset of the principal's actions that are consistent with agent i's observation, his posterior belief only attaches strictly positive probability to actions that are optimal within this subset. This axiom has no bite when each agent is abused with positive probability. When some agents are abused with zero probability, Axiom 3's requirement on off-path belief is similar to that in proper equilibrium (Myerson 1978). For our proofs to go through, one can replace Axiom 3 with the following Markovian axiom, that for every $\boldsymbol{a}$ and $\boldsymbol{a'}$ such that the evaluator attaches the same probability with which the principal is guilty, then $q(\boldsymbol{a}) = q(\boldsymbol{a'})$.

First, we establish the existence of equilibrium in the two-agent scenario that satisfies these axioms.

Proposition 3'. *There exists $\overline{L} > 0$ such that when $L > \overline{L}$, there exists a sequential equilibrium that satisfies Axioms 1, 2, and 3.*

Next, we show that all sequential equilibria that survive these refinements are symmetric:

Theorem 1'. *There exists $\overline{L} \in \mathbb{R}_+$ such that when $L > \overline{L}$, in every sequential equilibrium that satisfies Axioms 1, 2, and 3, there exists a triple $(\omega_m^*, \omega_m^{**}, \pi_m)$ such that:*

1. For every $i \in \{1,2\}$, agent i reports when $\left\{\omega_i \leq \omega_m^*$ and $\theta_i = 0\right\}$ or $\left\{\omega_i \leq \omega_m^{**}$ and $\theta_i = 1\right\}$.

[24] We have omitted the evaluator's belief and each agent's belief about other agents' ω and whether other agents are strategic or mechanical. The former can be computed via Bayes Rule. The latter does not depend on that agent's information set under the "*no-signaling what you don't know*" condition, which is satisfied in all sequential equilibria.

2. *The principal chooses* $(\theta_1, \theta_2) = (1, 1)$ *with probability* $1 - \pi_m$, $(\theta_1, \theta_2) = (1, 0)$ *with probability* $\pi_m/2$, *and* $(\theta_1, \theta_2) = (0, 1)$ *with probability* $\pi_m/2$.

The proof is in Online Appendix B. According to Theorem 1', every equilibrium that survives our refinement must be symmetric, both in the agents' strategies, and in the principal's strategy. This leads to the following implications. First, since each agent is abused with positive probability, his belief can be pinned down via Bayes rule. Second, Theorem 1 and 1' together imply that all sequential equilibria that satisfy Axioms 1, 2 and 3 also possess the properties stated in Theorems 1 and 2, namely, the principal abuses at most one agent, the probability of crime increases and the informativeness of report decreases compared to the single-agent benchmark. Moreover, as $L \to \infty$, the probability of crime converges to π^* and the informativeness of report converges to 1.

B Proofs of Main Results

B.1 Proof of Theorem 1, Statement 4

First, we show that $\omega_m^* > \omega_s^*$. Suppose towards a contradiction that $\omega_m^* \leq \omega_s^*$, then the comparison between (3.1) and (3.8) implies that

$$q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big) \leq q_s.$$

Therefore,

$$\begin{aligned} q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_s^*) - \Phi(\omega_s^{**})\Big) &\leq q_s\Big(\Phi(\omega_s^*) - \Phi(\omega_s^{**})\Big) \\ = 1/\delta L = q_m\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big)&\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big). \end{aligned} \tag{B.1}$$

Since $\omega_m^* - \omega_m^{**} < b = \omega_s^* - \omega_s^{**}$ and $\omega_m^* < \omega_s^*$, we have:

$$\Phi(\omega_m^*) - \Phi(\omega_m^{**}) < \Phi(\omega_s^*) - \Phi(\omega_s^{**}). \tag{B.2}$$

Inequality (B.2) implies that

$$\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_s^*) - \Phi(\omega_s^{**})\Big) > \Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big), \tag{B.3}$$

which contradicts (B.1). This implies that $\omega_m^* > \omega_s^*$. Moreover, according to Lemma 3.3,

$$0 < \omega_m^* - \omega_m^{**} < b = \omega_s^* - \omega_s^{**},$$

we know that $\omega_m^* > \omega_s^*$ implies $\omega_m^{**} > \omega_s^{**}$. The comparison between $\mathcal{I}_s$ and $\mathcal{I}_m$ immediately follows. This is because $\omega_m^{**} > \omega_s^{**}$ and $\omega_m^* - \omega_m^{**} < \omega_s^* - \omega_s^{**}$ imply that $\mathcal{I}_s > \mathcal{I}_m$. The comparison between π_s and π_m can then be obtained by comparing (3.6) to (3.15), which yields $\pi_s < \pi_m$.

Next, we show $q_m > q_s$. Given that $\omega_m^* > \omega_s^*$, then the comparison between (3.1) and (3.8) implies that $q_m Q_0 > q_s$. That is $1 \geq Q_0 > q_s/q_m$, which implies that $q_m > q_s$.

B.2 Principal's Incentives: Strategic Substitutes or Strategic Complements

We establish the complementarity or substitutability between the principal's choices of θ_1 and θ_2.

Lemma B.1. *In every equilibrium that satisfies Axiom 1, the principal's choices of* θ_1 *and* θ_2 *are strategic substitutes if the value of (3.7) is strictly positive and are strategic complements if the value of (3.7) is strictly negative.*

Proof. First, let us fix $\theta_2 = 1$, by changing θ_1 from 1 to 0, the principal increases the probability of conviction by:

$$(\Psi_1^* - \Psi_1^{**})\Big((1 - \Psi_2^{**})\big(q(1,0) - q(0,0)\big) + \Psi_2^{**}\big(q(1,1) - q(0,1)\big)\Big).$$

Similarly, fix $\theta_2 = 0$, by changing θ_1 from 1 to 0, the principal increases the probability of conviction by:

$$(\Psi_1^* - \Psi_1^{**})\Big((1 - \Psi_2^{*})\big(q(1,0) - q(0,0)\big) + \Psi_2^{*}\big(q(1,1) - q(0,1)\big)\Big).$$

The first expression is greater than the second one, or equivalently, the principal's choices of θ_1 and θ_2 are strategic complements, if and only if:

$$(\Psi_1^* - \Psi_1^{**})(\Psi_2^* - \Psi_2^{**})\Big(q(1,0) + q(0,1) - q(0,0) - q(1,1)\Big) > 0.$$

Since $\omega_i^* > \omega_i^{**}$ for $i \in \{1,2\}$, we know that $(\Psi_1^* - \Psi_1^{**})(\Psi_2^* - \Psi_2^{**}) > 0$. Therefore, the above inequality is equivalent to

$$q(1,0) + q(0,1) - q(0,0) - q(1,1) > 0,$$

which concludes the proof of Lemma B.1. □

B.3 Proof of Theorem 2

According to Lemma 3.3 and Lemma 3.4, we only need to show that $\omega_m^* \to -\infty$ or $\omega_m^{**} \to -\infty$ as $L \to \infty$. Recall that the principal's indifference condition is given by:

$$(\delta L)^{-1} = q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big) \tag{B.4}$$

Suppose towards a contradiction that there exist $\{L(n)\}_{n=1}^{\infty}$ and $\Big\{\omega_m^*(n), \omega_m^{**}(n), q_m(n), \pi_m(n)\Big\}_{n=1}^{\infty}$ such that:

1. $L(n) \geq \overline{L}$ for every $n \in \mathbb{N}$, and $\lim_{n\to\infty} L(n) = \infty$;
2. $(\omega_m^*(n), \omega_m^{**}(n), q_m(n), \pi_m(n))$ is an equilibrium when $L = L(n)$;
3. $\lim_{n\to\infty} \omega_m^{**}(n) = \omega^{**}$ for some $\omega^{**} \in \mathbb{R}$.

Since $\delta\Phi(\omega_m^{**}(n)) + (1-\delta)\alpha$ is bounded away from 0, (B.4) implies that at least one of the following statements is true:

1. *either* there exists a subsequence $\{k_n\}_{n=1}^{\infty} \subset \mathbb{N}$ such that: $\lim_{n\to\infty} q_m(k_n) = 0$.
2. *or* there exists a subsequence $\{k_n\}_{n=1}^{\infty} \subset \mathbb{N}$ such that: $\lim_{n\to\infty}\big(\Phi(\omega_m^*(k_n)) - \Phi(\omega_m^{**}(k_n))\big) = 0$. According to requirement 3, this is equivalent to $\lim_{n\to\infty}\big(\omega_m^*(k_n) - \omega_m^{**}(k_n)\big) = 0$.

First, suppose that $\lim_{n\to\infty} q_m(k_n) = 0$ for some subsequence $\{k_n\}_{n=1}^{\infty}$. Then (3.8) and (3.9) imply that both $\omega_m^*(k_n)$ and $\omega_m^{**}(k_n)$ converge to $-\infty$. This contradicts the third requirement that the sequence $\omega_m^{**}(n)$ converges to some finite number ω^{**}.

Next, suppose that $\lim_{n\to\infty} \omega_m^*(k_n) - \omega_m^{**}(k_n) = 0$ for some subsequence $\{k_n\}_{n=1}^{\infty}$. Since $\omega_m^{**}(k_n)$ converges to an interior number, both $Q_1(k_n)$ and $Q_0(k_n)$ are bounded away from 0, which suggests that $Q_1(k_n)/Q_0(k_n)$ converges to 1. From the previous step, we know that there exists no subsequence of $\{k_n\}_{n=1}^{\infty}$ such that $q_m(k_n)$ converges to 0. That is to say, there exists $\eta > 0$ such that $q_m(k_n) \geq \eta$ for every $n \in \mathbb{N}$. Expression (3.13) then suggests that $\omega_m^*(k_n) - \omega_m^{**}(k_n)$ converges to b. This contradicts the hypothesis that $\lim_{n\to\infty} \omega_m^*(k_n) - \omega_m^{**}(k_n) = 0$.

B.4 Proof of Proposition 7

We start from deriving formulas for the agents' reporting cutoffs $(\omega_n^*, \omega_n^{**})$, the informativeness of reports $\mathcal{I}_n$ and the equilibrium probability of crime π_n. In an n-agent economy and for every $i \in \{1, 2, ..., n\}$, agent i's reporting cutoff when $\theta_i = 0$ is:

$$\omega_n^* = c - \frac{c}{q_n Q_{0,n}}. \tag{B.5}$$

His reporting cutoff when $\theta_i = 1$ is:

$$\omega_n^{**} = -b + c - \frac{c}{q_n Q_{1,n}}, \tag{B.6}$$

in which

$$Q_{0,n} \equiv \Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-1} \tag{B.7}$$

and

$$Q_{1,n} \equiv \frac{n\mathcal{I}_n}{(n-1)l^* + n\mathcal{I}_n}\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-1}$$
$$+ \frac{(n-1)l^*}{(n-1)l^* + n\mathcal{I}_n}\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-2}\Big(\delta\Phi(\omega_n^{*}) + (1-\delta)\alpha\Big). \tag{B.8}$$

In symmetric unanimous equilibria, the aggregate informativeness of reports is given by the ratio between the probability with which n agents report conditional on $\Pi_{i=1}^n \theta_i = 0$ and the probability with which n agents report conditional on $\Pi_{i=1}^n \theta_i = 1$. This leads to the following formula:

$$\mathcal{I}_n = \frac{\delta\Phi(\omega_n^*) + (1-\delta)\alpha}{\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha}.$$

Since the evaluator is indifferent between convicting and acquitting the principal when there are n reports, we have:

$$\mathcal{I}_n = \frac{\pi^*}{1-\pi^*} \Big/ \frac{\pi_n}{1-\pi_n}. \tag{B.9}$$

When L is large enough, the principal is indifferent between abusing one agent and abusing no agent, which leads to the indifference condition:

$$\frac{1}{\delta L} = \delta q_n \Big(\Phi(\omega_n^*) - \Phi(\omega_n^{**})\Big)\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-1}. \tag{B.10}$$

Reporting Cutoffs & Distance Between Cutoffs: In this part, we show that $\omega_k^* > \omega_n^*$. Suppose towards a contradiction that $\omega_k^* \leq \omega_n^*$, then according to (B.5), we have:

$$q_k\Big(\delta\Phi(\omega_k^{**}) + (1-\delta)\alpha\Big)^{k-1} \leq q_n\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-1}. \tag{B.11}$$

Therefore, $q_k Q_{0,k} \leq q_n Q_{0,n}$ which is equivalent to:

$$q_k\Big(\delta\Phi(\omega_k^{**}) + (1-\delta)\alpha\Big)^{k-1}\Big(\Phi(\omega_n^*) - \Phi(\omega_n^{**})\Big) \leq q_n\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-1}\Big(\Phi(\omega_n^*) - \Phi(\omega_n^{**})\Big)$$
$$= q_k\Big(\delta\Phi(\omega_k^{**}) + (1-\delta)\alpha\Big)^{k-1}\Big(\Phi(\omega_k^*) - \Phi(\omega_k^{**})\Big).$$

This implies that

$$\Phi(\omega_n^*) - \Phi(\omega_n^{**}) \leq \Phi(\omega_k^*) - \Phi(\omega_k^{**}). \tag{B.12}$$

Since $\omega_k^* \leq \omega_n^*$, (B.12) can only be true when

$$\omega_n^* - \omega_n^{**} \leq \omega_k^* - \omega_k^{**}, \tag{B.13}$$

which in turn implies that $\omega_k^{**} \leq \omega_n^{**}$ and therefore $q_k Q_{1,k} \leq q_n Q_{1,n}$. Computing the two sides of (B.13) by subtracting (B.6) from (B.5), we have:

$$\omega_n^* - \omega_n^{**} = b - \frac{c}{q_n} \frac{Q_{1,n} - Q_{0,n}}{Q_{1,n} Q_{0,n}} \text{ and } \omega_k^* - \omega_k^{**} = b - \frac{c}{q_k} \frac{Q_{1,k} - Q_{0,k}}{Q_{1,k} Q_{0,k}}.$$

Due to the previous conclusion that $q_k Q_{0,k} \leq q_n Q_{0,n}$ and $q_k Q_{1,k} \leq q_n Q_{1,n}$, (B.13) is true only when

$$q_n(Q_{1,n} - Q_{0,n}) \geq q_k(Q_{1,k} - Q_{0,k}). \tag{B.14}$$

Since

$$Q_{1,n} - Q_{0,n} = \frac{(n-1)l^*}{(n-1)l^* + n\mathcal{I}_n} \delta\Big(\Phi(\omega_n^*) - \Phi(\omega_n^{**})\Big)\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-2}$$

and the term

$$\delta\Big(\Phi(\omega_n^*) - \Phi(\omega_n^{**})\Big)\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-2} = L^{-1} q_n^{-1} \frac{1}{\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha}$$

according to (B.10), we know that (B.14) is equivalent to:

$$\frac{(n-1)l^*}{(n-1)l^*\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big) + n\Big(\delta\Phi(\omega_n^*) + (1-\delta)\alpha\Big)}$$

$$\geq \frac{(k-1)l^*}{(k-1)l^*\Big(\delta\Phi(\omega_k^{**}) + (1-\delta)\alpha\Big) + k\Big(\delta\Phi(\omega_k^*) + (1-\delta)\alpha\Big)}$$

which in turn reduces to:

$$(n-1)(k-1)l^*\Big(\delta\Phi(\omega_k^{**}) + (1-\delta)\alpha\Big) + (n-1)k\Big(\delta\Phi(\omega_k^*) + (1-\delta)\alpha\Big)$$

$$\geq (n-1)(k-1)l^*\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big) + (k-1)n\Big(\delta\Phi(\omega_n^*) + (1-\delta)\alpha\Big)$$

The above inequality cannot be true since $\delta\Phi(\omega_k^{**}) + (1-\delta)\alpha < \delta\Phi(\omega_n^{**}) + (1-\delta)\alpha$, $\delta\Phi(\omega_k^*) + (1-\delta)\alpha < \delta\Phi(\omega_n^*) + (1-\delta)\alpha$ and $(n-1)k < (k-1)n$. The last inequality holds due to the assumption that $k > n$. This leads to a contradiction which shows that $\omega_k^* > \omega_n^*$ whenever $k > n$.

Notice that up until the last step, we did not use the fact that $k > n$. Given the previous conclusion that $\omega_k^* > \omega_n^*$ and repeat the same reasoning up until (B.13), we know that

$$\omega_n^* - \omega_n^{**} > \omega_k^* - \omega_k^{**}, \tag{B.15}$$

and this further implies that $\omega_k^{**} > \omega_n^{**}$.

Informativeness & Probability of Crime: In this part, we establish the comparison between informativeness by showing that $\mathcal{I}_n > \mathcal{I}_k$, that is, having more agents decreases the net informativeness of reports. Due to the one-to-one mapping between net informativeness and the probability of at least one assault taking place, this also implies that $\pi_k > \pi_n$, that is, the probability of crime increases.

Applying (B.5) and (B.6) to both n and k, we obtain the following expression for the ratios:

$$\frac{\omega_n^* - c}{\omega_k^* - c} = \frac{q_k Q_{0,k}}{q_n Q_{0,n}} \quad \text{and} \quad \frac{\omega_n^{**} + b - c}{\omega_k^{**} + b - c} = \frac{q_k Q_{0,k}(\beta_k + (1-\beta_k)\mathcal{I}_k)}{q_n Q_{0,n}(\beta_n + (1-\beta_n)\mathcal{I}_n)}. \tag{B.16}$$

First, we show that

$$\frac{\omega_n^* - c}{\omega_k^* - c} > \frac{\omega_n^{**} + b - c}{\omega_k^{**} + b - c}. \tag{B.17}$$

Suppose towards a contradiction that the opposite of (B.17) is true, then

$$\frac{\omega_n^{**}+b-c-(\omega_n^*-c)}{\omega_k^{**}+b-c-(\omega_k^*-c)} \geq \frac{\omega_n^*-c}{\omega_k^*-c}. \tag{B.18}$$

The RHS of (B.18) is strictly greater than 1 since $0 > \omega_k^* > \omega_n^*$. The LHS of (B.18) being greater than 1 is equivalent to

$$b-(\omega_n^*-\omega_n^{**}) > b-(\omega_k^*-\omega_k^{**})$$

which contradicts the previous conclusion in (B.15). This establishes (B.17). This together with (B.16) imply that

$$\beta_k+(1-\beta_k)\mathcal{I}_k < \beta_n+(1-\beta_n)\mathcal{I}_n.$$

Plugging in the expressions of $\mathcal{I}_n$ and $\mathcal{I}_k$ in (B.9), we have:

$$\mathcal{I}_k\big(k+(k-1)l^*\big)\big(n\mathcal{I}_n+(n-1)l^*\big) < \mathcal{I}_n\big(n+(n-1)l^*\big)\big(k\mathcal{I}_k+(k-1)l^*\big).$$

Let $\Delta \equiv \mathcal{I}_k - \mathcal{I}_n$, the above inequality reduces to:

$$(k-n)\mathcal{I}_n(\mathcal{I}_n+\Delta-1) < k\Delta - \Big(l^*(k-1)(n-1)+nk\Big)\Delta.$$

Suppose towards a contradiction that $\Delta \geq 0$, then the LHS is strictly positive since $\mathcal{I} > 1$ and $k > n$. The RHS is negative since $l^*(k-1)(n-1)+nk > k$. This leads to a contradiction which implies that $\Delta < 0$ and therefore, $\mathcal{I}_n > \mathcal{I}_k$.

C Proofs in Single-Agent Benchmark

Proof of Proposition 1: The proof consists of three parts.

Necessity: When (3.4) fails, then the principal's expected cost of committing a crime in any equilibrium is at most:

$$\delta q_s L\Big(\Phi(\omega_s^*)-\Phi(\omega_s^{**})\Big) \leq \delta L\big(\Phi(0)-\Phi(-b)\big). \tag{C.1}$$

The maximum on the RHS is attained when $q_s = 1$, $\omega_s^* = 0$ and $\omega_s^{**} = -b$. The value of the RHS is strictly less than his benefit from committing a crime, which is 1. This leads to a contradiction.

Sufficiency: When (3.4) holds, then for a fixed c, notice that ω_s^* and ω_s^{**} are strictly increasing in q_s. Since $\omega_s^*, \omega_s^{**} \leq 0$, the distance between ω_s^* and ω_s^{**} is b and the pdf of $\mathcal{N}(\mu,\sigma^2)$ is strictly increasing in ω when $\omega < 0$, we know that $\Phi(\omega_s^*)-\Phi(\omega_s^{**})$ is also strictly increasing in q_s. This implies that the LHS of (3.5), which is the cost of abusing the agent, is strictly increasing in q_s. Inequality (3.4) ensures the existence and uniqueness of $(\omega_s^*, \omega_s^{**}, q_s)$ that solves (3.1), (3.2) and (3.5) since the LHS of (3.5) is strictly less than $1/L$ when $q_s = 0$ and is weakly more than $1/L$ when $q_s = 1$.

Uniqueness: When (3.4) holds with strictly inequality, the equilibrium level of q_s is interior. As a result, the probability with which the principal is guilty according to the evaluator's posterior belief is π^*. As a result, π_s is uniquely pinned down by (3.6). □

Proof of Proposition 2: When $L \to \infty$ while holding c constant, the RHS of (3.5) converges to 0. Suppose towards a contradiction that q_s converges to some strictly positive number $\underline{q}$ along some sequence $\{L_n\}_{n=1}^{\infty}$

with $\lim_{n\to\infty} L_n = \infty$. Then ω_s^* and ω_s^{**} converge to $-c(1-\underline{q})/\underline{q}$ and $-b-c(1-\underline{q})/\underline{q}$, respectively. The LHS of (3.5) converges to

$$\delta\underline{q}\Big(\Phi(-c\frac{1-\underline{q}}{\underline{q}}) - \Phi(-b-c\frac{1-\underline{q}}{\underline{q}})\Big) \tag{C.2}$$

which is strictly bounded away from 0. This leads to a contradiction and establishes that $q_s \to 0$. The expressions for ω_s^* and ω_s^{**} in (3.1) and (3.2) imply that both cutoffs converge to $-\infty$. In the limiting economy in which $\delta \to 1$, we have:

$$\lim_{\omega_s^*\to-\infty}\lim_{\delta\to 1}\frac{\delta\Phi(\omega_s^*)+(1-\delta)\alpha}{\delta\Phi(\omega_s^*-b)+(1-\delta)\alpha} = \infty. \tag{C.3}$$

The above equation makes use of the observation that the tail events of normal distributions are arbitrarily informative, or formally, $\lim_{\omega\to-\infty}\Phi(\omega)/\Phi(\omega-b) \to \infty$ for every $b > 0$. That is to say, it extends to all distributions of ω with thin left tails. As a result, the equilibrium probability of crime π_s vanishes to 0. □

References

[1] Ali, S. Nageeb, Maximilian Mihm and Lucas Siga (2018) "Adverse Selection in Distributive Politics," Working Paper.

[2] Austen-Smith, David and Jeffrey Banks (1996) "Information Aggregation, Rationality, and the Condorcet Jury Theorem," *American Political Science Review*, 90(1), 34-45.

[3] Baliga, Sandeep, Ethan Bueno de Mesquita and Alexander Wolitzky (2019) "Deterrence with Imperfect Attribution," Working Paper.

[4] Baliga, Sandeep and Tomas Sjöström (2004) "Arms Races and Negotiations," *Review of Economic Studies*, 71(2), 351-369.

[5] Banerjee, Abhijit (1992) "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107(3), 797-817.

[6] Becker, Gary (1968) "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 76(2), 169-217.

[7] Bhattacharya, Sourav (2013) "Preference Monotonicity and Information Aggregation in Elections," *Econometrica*, 81(3), 1229-1247.

[8] Bikhchandani, Sushil, David Hirshleifer, Ivo Welch (1992) "A Theory of Fads , Fashion , Custom , and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100(5), 992-1026.

[9] Carlson, Hans and Eric Van Damme (1993) "Global Games and Equilibrium Selection," *Econometrica*, 61(5), 989-1018.

[10] Chamberlin, John (1974) "Provision of Collective Goods As a Function of Group Size," *The American Political Science Review*, 68(2), 707-716.

[11] Chassang, Sylvain and Gerard Padró i Miquel (2010) "Conflict and Deterrence under Strategic Risk," *Quarterly Journal of Economics*, 125(4), 1821-1858.

[12] Chassang, Sylvain and Gerard Padró i Miquel (2018) "Corruption, Intimidation and Whistle-Blowing: A Theory of Inference from Unverifiable Reports," NBER working paper.

[13] Crémer, Jacques and Richard McLean (1985) "Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist when Demands are Interdependent," *Econometrica*, 53(2), 345-361.

[14] Crémer, Jacques and Richard McLean (1988) "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica*, 56(6), 1247-1257.

[15] Dresher, Melvin (1962) "A Sampling Inspection Problem in Arms Control Agreements – A Game-Theoretic Analysis," Memorandum, The RAND Corporation, Santa Monica, California.

[16] Feddersen, Timothy and Wolfgang Pesendorfer (1996) "The Swing Voter's Curse," *American Economic Review*, 86(3), 408-424.

[17] Fudenberg, Drew and David Levine (1995) "The Theory of Learning in Games," MIT Press.

[18] Grossman, Gene and Michael Katz (1983) "Plea Bargaining and Social Welfare," *American Economic Review*, 73(4), 749-757.

[19] Harris, John (1970) "On the Economics of Law and Order," *Journal of Political Economy*, 78(1), 165-174.

[20] Kaplow, Louis (2011) "On the Optimal Burden of Proof," *Journal of Political Economy*, 119(6), 1104-1140.

[21] Klement, Alon, and Zvika Neeman (2005) "Against Compromise: A Mechanism Design Approach," *Journal of Law, Economics, and Organization*, 21(2), 285-314.

[22] Landes, William (1971) "An Economic Analysis of the Courts," *Journal of Law and Economics*, 14(1), 61-107.

[23] Lee, Frances Xu and Wing Suen (2018) "Credibility of Crime Allegations," Working Paper.

[24] Miceli, Thomas (1991) "Optimal Criminal Procedure: Fairness and Deterrence," *International Review of Law and Economics* 11(1), 3-10.

[25] Morgan, John and Phillip Stocken (2008) "Information Aggregation in Polls," *American Economic Review*, 98(3), 864-896.

[26] Morris, Stephen and Hyun Song Shin (1998) "Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks," *American Economic Review*, 88(3), 587-597.

[27] Myerson, Roger (1978) "Refinements of the Nash Equilibrium Concept," *International Journal of Game Theory*, 7(2), 73-80.

[28] Ottaviani, Marco and Peter Norman Sørensen (2000) "Herd Behavior and Investment: Comment," *American Economic Review*, 90(3), 695-704.

[29] RAND Cooperation (2018) "Sexual Assault and Sexual Harassment in the US Military," Technical Report.

[30] Reinganum, Jennifer (1988) "Plea Bargaining and Prosecutorial Discretion," *American Economic Review*, 78(4), 713-728.

[31] Scharfstein, David and Jeremy Stein (1990) "Herd Behavior and Investment," *Amercian Economic Review*, 80(3), 465-479.

[32] Siegel, Ron and Bruno Strulovici (2018) "Judicial Mechanism Design," Working Paper.

[33] Silva, Francesco (2018) "If We Confess Our Sins," Working Paper.

[34] Smith, Lones and Peter Norman Sørensen (2000) "Pathological Outcomes of Observational Learning," *Econometrica*, 68(2), 371-398.

[35] Stigler, George (1970) "The Optimal Enforcement of Laws," *Journal of Political Economy*, 78(3), 526-536.

[36] Strulovici, Bruno (2018) "Can Society Learn from Anethical Agents? A Theory of Mediated Learning with Information Attrition," Working Paper.

[37] USMSPB (2018) "Update on Sexual Harassment in the Federal Workplace," Research Brief.