The Use and Misuse of Coordinated Punishments

Daniel Barron and Yingni Guo*

November 19, 2019

Abstract

Communication facilitates cooperation by ensuring that deviators are collectively punished. We explore how players might misuse messages to threaten one another, and we identify ways in which organizations can deter misuse and restore cooperation. In our model, a principal plays trust games with a sequence of short-run agents who communicate with one another. An agent can shirk and then extort pay by threatening to report that the principal deviated. We show that these threats can completely destroy cooperation. Investigations of agents' efforts, or dyadic relationships between the principal and each agent, can deter extortion and restore some cooperation. Investigations of the principal's action, on the other hand, typically don't help. Our analysis suggests that collective punishments improve cooperation only if they are designed with an eye towards discouraging misuse.

[*] Barron: Northwestern University, Kellogg School of Management, Evanston IL 60208; email: d-barron@kellogg.northwestern.edu. Guo: Northwestern University, Economics Department, Evanston IL 60208; email: yingni.guo@northwestern.edu. The authors would like to thank Nageeb Ali, Charles Angelucci, Nemanja Antic, Alessandro Bonatti, Renee Bowen, Joyee Deb, Wouter Dessein, Matthias Fahn, Benjamin Friedrich, George Georgiadis, Marina Halac, Peter Klibanov, Ilan Kremer, Nicolas Lambert, Stephan Lauermann, Jin Li, Elliot Lipnowski, Shuo Liu, Bentley MacLeod, David Miller, Joshua Mollner, Dilip Mookherjee, Arijit Mukherjee, Jacopo Perego, Michael Powell, Luis Rayo, Jonah Rockoff, Mark Satterthwaite, Andy Skrzypacz, Takuo Sugaya, Jeroen Swinkels, Joel Watson, and audiences at many conferences, workshops, and seminars.We thank the UCSD theory reading group for comments on a draft of this paper, and Andres Espitia for excellent research assistance.

1 Introduction

Productive relationships thrive on the enthusiastic cooperation of their participants. In many settings, individuals cooperate because they expect opportunistic behavior to be punished (Malcomson (2013)). Communication plays an essential role in coordinating these punishments, since it allows those who do not directly observe misbehavior to nevertheless punish the perpetrator. These coordinated punishments are central to relationships between managers and workers (Levin (2002)), suppliers and their customers (Greif et al. (1994); Bernstein (2015)), community members (Ostrom (1990)), and participants in online marketplaces (Hörner and Lambert (2018)).

Once armed with the power to trigger coordinated punishments, individuals face a grave temptation: they can extort concessions from their partners by threatening to *falsely* report opportunistic behavior (Gambetta (1993); Dixit (2003a, 2007)). In this paper, we explore how individuals might misuse coordinated punishments. We emphasize two overarching takeaways. First, we show that misuse is a serious problem that can completely destroy cooperation. Second, we identify practical ways that organizations can restore cooperation in the face of such misuse.

To illustrate the use and misuse of coordinated punishments, consider a manager who wants to motivate her workers to exert effort beyond their narrowly contracted duties. Workers are willing to strive for excellent performance only if they trust their manager to reward their efforts (Gibbons and Henderson (2013)). In this context, an institution that allows workers to collectively punish misbehaving managers, such as a labor union (Freeman and Medoff (1979)) or a job review platform like Glassdoor.com, can deter managers' opportunistic behavior and potentially encourage highly productive effort.

We argue that unless such institutions are carefully designed, workers face the temptation to subvert them in pursuit of private gain. For example, in the 1980s, workers at General Motors' Fremont plant used the threat of accumulated grievances to get away with "shirking" activities, including absenteeism and drug use on the plant floor, that neither management nor the union condoned. These workers misused the threat of coordinated punishments – in the form of labor unrest – to demand undeserved compensation. The result was fractious manager-worker relationships, low productivity, and the eventual closure of the plant (Glass and Langfitt (2015)).

The use of coordinated punishments, and the potential for misuse, extend far beyond the factory floor. Firms use industry associations like the Financial Industry Regulatory Authority (FINRA) to share information about misbehaving employees. During its recent scandal, however, Wells Fargo faced allegations that it punished employees who spoke up about fraudulent practices by falsely reporting them for unethical behavior (Arnold and Smith (2016)). Similar misuse is a real concern in online marketplaces. In the early days of eBay, for example, sellers extorted positive reviews from buyers by threatening to reciprocate on any negative review, undermining sellers' incentives to exert effort and leading to less satisfied buyers (Klein et al. (2016)).[1]

To explore the use and misuse of coordinated punishments, we consider a model of a long-run principal who interacts with a sequence of short-run agents. Each agent exerts costly effort to benefit the principal, who can then choose to pay him. Agents observe only their own interactions but can communicate with one another. To capture the idea that extortion entails action-contingent threats – i.e., "pay me *or else* I will punish you" – we allow each agent to make a **threat** when he chooses his effort. This threat, which is observed by the principal but not by other agents, associates a message to each possible payment. Agents then follow through on their threats.

In this model, misuse completely destroys cooperation. The principal is willing to pay an agent only if she would otherwise be punished by future agents. Communication is therefore essential for cooperation. Once endowed with messages that trigger punishments, however, an agent can extort the principal by shirking and then threatening to trigger punishments

[1] Platforms have policies to combat these types of extortionary threats. See, for instance, TripAdvisor's policy at https://www.tripadvisor.com/TripAdvisorInsights/w592. Despite these policies, extortion remains a problem. See Conti (2019) for an example on AirBnB.

unless the principal pays him. Since this threat is enough to induce the principal to pay a hard-working agent, it is also enough to induce her to pay a shirking agent. Thus, the pay that an agent can demand is essentially independent of his effort. The stark implication of this logic is that agents do not exert any effort.

After establishing this impossibility result, we explore how organizations can deter extortion and encourage cooperation. We focus on two instruments that are available in many cooperative endeavors: *investigations*, which we model as public signals of either the agents' efforts or the principal's transfers, and *ongoing dyadic relationships*, which we model as a coordination game played by the principal and each agent.

The unifying idea of these instruments is that agents are willing to exert effort only if doing so affects how severely they can threaten the principal. Define an agent's **leverage** over the principal as the harshest punishment that he can trigger with his report. Each agent can extort any transfer that is smaller than his leverage. If an agent's leverage is necessarily independent of his effort, as it is in our baseline model, then he has no incentive to exert effort. If an agent's leverage is increasing in his effort, on the other hand, then he might exert effort in order to increase his leverage, so that he can demand higher pay. An instrument is valuable exactly when it can tie leverage to effort in this way.

Building on this idea, we first show that investigations of agents' efforts typically improve cooperation, whereas investigations of the principal's transfers typically do not. Effort signals are useful for deterring extortion, not because agents are directly rewarded or punished on the basis of these signals, but instead because these signals can tie leverage to effort. They do so by ensuring that harder-working agents can trigger harsher coordinated punishments. Agents are then willing to exert effort in order to obtain higher leverage and demand higher pay. In contrast, transfer signals can reveal whether the principal paid an agent but not whether that pay was *deserved* or not. Hence, such signals typically cannot tie leverage to effort. The only exception is that, under stringent conditions, transfer signals can make the principal indifferent between transfers. But even under these stringent conditions, the extent of cooperation is limited by the need for occasional on-path punishments.

We then study how ongoing relationships between the principal and each agent can deter extortion. These dyadic relationships potentially expand the scope for coordinated punishments (Levin (2002)); unless extortion is deterred, however, stronger coordinated punishments simply lead to more lucrative extortion opportunities. As with investigations, dyadic relationships deter extortion by tying leverage to effort. To do so, these relationships reward the principal for *refusing* to pay a shirking agent, which limits that agent's leverage. Dyadic relationships therefore complement coordinated punishments: agents with strong dyadic relationships can be given lots of leverage without opening the door to extortion, while agents with weak dyadic relationships are optimally given minimal access to coordinated punishments.

The premise of our analysis is that, while organizations can potentially benefit from coordinated punishments, they cannot perfectly control how their members actually use these punishments. Successful organizations must therefore focus at least as much on deterring misuse of coordinated punishments as on optimizing their proper use, which demands a fundamentally different approach to designing incentive systems. In our setting, these systems are embedded in the rules or culture of the organization, as represented by an equilibrium of our dynamic game. But our lessons extend to other settings in which cheap-talk reports are used to motivate prosocial behavior.

Related Literature

Our contribution is to explore how extortion undermines coordinated punishments and how organizations can combat it. Therefore, we build on the literature that studies how coordinated punishments support cooperation (Milgrom et al. (1990), Greif et al. (1994), Dixit (2003a,b)). Much of this literature focuses on networks of players and has as its goal the identification of network structures or equilibrium strategies that are particularly conducive to cooperation (Lippert and Spagnolo (2011), Wolitzky (2013), Ali and Miller (2013, 2016), Ali et al. (2017)). Especially related is Ali and Miller (2016), which shows that players might not report deviations if doing so reveals that they are more willing to renege on their own promises. Extortion is a different but complementary obstacle to coordinated punishments.

Since extortion is inherently action-contingent – i.e., "pay me *or else* I will punish you" – our analysis is related to a growing literature on action-contingent threats and promises. Like us, some of these papers assume players commit to threats in order to allow for action-contingent deviations (Wolitzky (2012), Chassang and Padro i Miquel (2018), Ortner and Chassang (2018)).[2] In our setting, we can also re-interpret commitment as an equilibrium refinement of the game without commitment, which is related to the approach taken in Zhu (2018, 2019). We contribute to this literature by studying how misuse can destroy cooperation by undermining coordinated punishments and exploring new ways for organizations to deter misuse by tying leverage to effort.

Much of the literature on cooperation focuses on the use of coordinated punishments rather than the potential for misuse. Dixit (2003a, 2007) is perhaps the first to formally model the misuse of coordinated punishments, although those models study centralized enforcers rather than decentralized communication. Bowen et al. (2013), which studies local adaptation in communities, considers a type of misuse that differs from extortion in that it is not action-contingent. The literature on coalitional deviations in repeated games (Ali and Liu (2018), Liu (2019)) is more closely related, as extortion resembles a bilateral coalitional deviation in which the agents have the bargaining power. In contrast to those papers, however, agents make threats concurrently with effort, rather than at the start of the period, and the resulting actions (effort and transfers) are not publicly observed.

In our setting, an agent essentially threatens the principal with a bad "outside option" unless she pays him. Our paper is therefore connected to the literature on renegotiation and bargaining in repeated games. Particularly related are papers that allow players to bargain

[2] Indeed, Ortner and Chassang (2018) have an appendix that studies extortion. However, that appendix assumes that reports lead to exogenous and fixed punishments, while the point of our analysis is to show how to optimally link messages to punishments.

over surplus in equilibrium (Baker et al. (2002), Miller and Watson (2013), Halac (2012, 2015), Goldlucke and Kranz (2017), Miller et al. (2018)). By focusing on communication across agents, our paper studies a setting in which the principal's "outside option" depends on how messages affect future equilibrium play.

More broadly, our framework builds on the relational contracting literature (Bull (1987), MacLeod and Malcomson (1989), Baker et al. (1994), Levin (2003)), especially those papers that study coordinated punishments (e.g., Levin (2002)). We introduce extortion as a threat that undermines such punishments. Recent papers have explored relational contracts in the presence of limited transfers (Fong and Li (2017), Barron et al. (2018)), asymmetric information (Halac (2012), Malcomson (2016)), or both (Li et al. (2017), Lipnowski and Ramos (2017), Guo and Hörner (2018)). We focus on a monitoring friction – agents do not observe one another's relationships – which implies that cooperation relies on communication. Other papers that study relational contracts with bilateral monitoring, including Board (2011), Andrews and Barron (2016), and Barron and Powell (2018), do not allow agents to communicate. We complement these papers by identifying a reason why communication might be ineffective at sustaining cooperation.

2 Model

Our baseline model is the following **extortion game.** A long-run principal ("she") interacts with a sequence of short-run agents (each "he"). In each period $t \in \{0, 1, 2, ...\}$, the principal and agent t play a trust game: agent t exerts effort, then the principal observes that effort and pays him. While this interaction is observed only by the principal and agent t, agent t can send a public message at the end of period t. Our key assumption is that before transfers are paid, agent t makes a **threat***,* which is a mapping from the transfer he receives to the message he sends. This threat is observed by the principal but not by other agents.

Formally, the stage game in period t is:

1. Agent t chooses his effort $e_t \in \mathbb{R}_+$ and a threat $\mu_t : \mathbb{R} \to M$, where M is a large, finite message space.[3] Both e_t and μ_t are observed by the principal but not by any other agent.

2. The principal makes a transfer to agent t, $s_t \geq 0$, which is observed by agent t but not by other agents.[4]

3. The message $m_t = \mu_t(s_t)$ is realized and observed by all players.

The principal's period-t payoff and agent t's utility are $(e_t - s_t)$ and $(s_t - c(e_t))$, respectively, where $c(\cdot)$ is twice continuously differentiable, strictly increasing, strictly convex, and satisfies $c(0) = c'(0) = 0$. We assume that there exists a first-best effort, e^{FB}, such that $c'(e^{FB}) = 1$. The principal has discount factor $\delta \in [0, 1)$, with corresponding normalized discounted payoffs $\Pi_t = (1 - \delta) \sum_{t'=t}^{\infty} \delta^{t'-t}(e_{t'} - s_{t'})$. Players observe a public randomization device (notation for which is suppressed) in every step of the stage game.

The principal observes everything, while agents observe only their own interactions with the principal and all messages. Our solution concept is Perfect Bayesian Equilibrium.[5] Some of our results focus on principal-optimal equilibria, which maximize the principal's *ex ante* expected payoff among all equilibria.

We occasionally compare our results to a benchmark without extortion. Define the **no-extortion game** as identical to the extortion game, except that each agent t chooses m_t at the end of period t rather than being committed to μ_t. In the no-extortion game, agents cannot shirk and then make action-contingent threats, so they cannot misuse communication.

Our goals are to (i) show why agents have the incentive to misuse communication, and (ii) explore how organizations can deter misuse in equilibrium. In our motivating applica-

[3] The assumption that M is finite simplifies the proofs (by ensuring that various maxima and minima exist) but is not essential for the results.

[4] For almost all of our results, the assumption that agents do not pay the principal is without loss. The exception is Section 5; we allow agents to pay the principal in that section.

[5] See Watson (2017). We consider a Perfect Bayesian Equilibrium in order to specify how agents form beliefs over histories, but since those beliefs do not play an important role in our arguments, our results would extend to various restrictions on off-path beliefs.

tions, agents misuse coordinated punishments by deviating and then making pay-contingent threats. The threat, μ_t, is a transparent way to allow this type of misuse, one that is similar to the approaches taken in Dixit (2003a), Wolitzky (2012), Chassang and Padro i Miquel (2018), and Ortner and Chassang (2018). In Online Appendix B, we show that we can re-interpret commitment to μ_t as an equilibrium refinement of the no-extortion game, since all equilibria in the extortion game remain equilibria in the no-extortion game. Under this interpretation, our approach is similar to that taken in Dewatripont (1987), Tranaes (1998), and Zhu (2018, 2019).

In our introductory example of General Motors' Fremont plant, agents are workers who exert effort (e_t) catching mistakes or improving quality, while the principal is a manager who can reward such efforts (via s_t). The manager follows through on promised rewards because she fears grievances (m_t) that trigger widespread labor unrest. At GM-Fremont, workers engaged in a variety of shirking behaviors, secure in the knowledge that they could still demand pay by threatening to file grievances (μ_t). We will show that agents have a similar incentive to shirk and then extort a transfer in the extortion game. Of course, the real world is richer than our model: unions typically investigate grievances, and individual workers have ongoing relationships with the manager. Sections 4 and 5 study how these instruments can be used to (imperfectly) combat extortion.

Online Appendix C considers alternative communication structures, including models in which the principal can send messages or make threats, as well as ones in which agents can make repeated threats. In most of these variants, extortion continues to undermine cooperation. We also identify particular communication structures that can lead to cooperation in equilibrium, although these positive results typically come with substantial caveats.

3 Threats Undermine Equilibrium Cooperation

This section shows how coordinated punishments are used and misused in equilibrium. We first illustrate how coordinated punishments sustain cooperation in the no-extortion game. Then, we show that extortionary threats lead cooperation to completely unravel. This impossibility result demonstrates the economics of extortion and forms the foundation for the rest of the analysis.

Cooperation requires agents to communicate with one another, since without communication an agent would have no way to punish the principal for deviating. In the no-extortion game, this type of communication is enough to sustain cooperation.

Proposition 1 *In the no-extortion game, $e_t = e^*$ and $s_t = c(e^*)$ in each $t \geq 0$ of every principal-optimal equilibrium, where e^* equals the minimum of e^{FB} and the positive root of $c(e) = \delta e$.*

Proof: We first argue that total equilibrium surplus is at most $e^* - c(e^*)$. By definition of e^{FB}, equilibrium surplus is at most $e^{FB} - c(e^{FB})$. If $c(e^{FB}) \leq \delta e^{FB}$, then $e^* = e^{FB}$ and the result follows. If $c(e^{FB}) > \delta e^{FB}$, then let $\bar{\Pi}$ be the principal's maximum *ex ante* equilibrium payoff. In any period $t \geq 0$ of any equilibrium, $(1-\delta)s_t \leq \delta\bar{\Pi}$ and $s_t - c(e_t) \geq 0$ must hold, since otherwise the principal or agent t could profitably deviate from s_t or e_t, respectively. Therefore, $(1-\delta)c(e_t) \leq \delta\bar{\Pi}$. Let $\bar{e}$ be the effort that maximizes $e - c(e)$ among any effort that is attained in any period of any equilibrium. Then $(1-\delta)c(\bar{e}) \leq \delta\bar{\Pi} \leq \delta(\bar{e} - c(\bar{e}))$ and so $c(\bar{e}) \leq \delta\bar{e}$. We conclude that $\bar{e} \leq e^* < e^{FB}$, so equilibrium surplus is at most $e^* - c(e^*)$.

Consider the following strategy profile for each period $t \geq 0$: if $m_{t'} = C$ in all $t' < t$, then agent t chooses $e_t = e^*$; the principal chooses $s_t = c(e^*)$ if $e_t = e^*$ and $s_t = 0$ otherwise; and agent t chooses $m_t = C$ if neither player deviates and $m_t = D$ otherwise. If $m_{t'} \neq C$ in at least one $t' < t$, then $e_t = s_t = 0$ and $m_t = D$.

Once $m_{t'} \neq C$ in some $t' < t$, this strategy profile specifies the one-shot equilibrium and so players cannot profitably deviate. If $m_{t'} = C$ in all $t' < t$, then agent t has no profitable deviation because he earns 0 on-path and no more than 0 from deviating. The principal has no profitable deviation because $(1-\delta)s_t \leq \delta(e^* - c(e^*))$ is implied by $c(e^*) \leq \delta e^*$. This strategy is therefore an equilibrium. It is principal-optimal because it generates total surplus $e^* - c(e^*)$, which is the maximum equilibrium surplus, and it holds agents at their min-max payoffs. Moreover, every principal-optimal equilibrium gives the principal a payoff of $e^* - c(e^*)$ and so must entail $e_t = e^*$ in every period. ■

The proof of Proposition 1 relies on the following equilibrium construction. On the equilibrium path, each agent sends the message C if the principal pays him and D otherwise. Future agents min-max the principal if they observe the message D. A shirking agent sends a message that is independent of the principal's transfer, so the principal pays him nothing. The principal would rather pay a hard-working agent a transfer than be punished, and each agent would rather exert effort than shirk and forgo the transfer, so this construction can motivate effort.

Proposition 1 summarizes a core idea from much of the literature on coordinated punishments. The principal rewards a hard-working agent because this agent will otherwise send a report that triggers future punishments. Implicit in this construction, and in much of the literature on coordinated punishments, is the requirement that shirking agents do not make similar threats, so that the principal *refrains* from paying a shirking agent.

The extortion game allows shirking agents to make exactly this type of threat. Our next result, which serves as a baseline for the rest of our analysis, shows that these threats destroy cooperation, so that agents shirk in any equilibrium.

Proposition 2 *In the extortion game, every equilibrium entails $e_t = s_t = 0$ in every $t \geq 0$.*

Proof: Fix a message history $m^{t-1} = (m_0, m_1, ..., m_{t-1})$, and let

$$\bar{\Pi} = \max_{m \in M} \left\{ \mathbb{E}\left[\Pi_{t+1} | m^{t-1}, m_t = m \right] \right\}$$

be the principal's maximum continuation surplus in period $t+1$ onwards that can be induced by some message. We denote this message $m_t = C$. Let $\underline{\Pi}$ be the similarly-defined minimum continuation payoff, with corresponding message $m_t = D$.

Suppose that agent t chooses some $e_t > 0$. He is willing to do so only if $s_t \geq c(e_t)$; the principal is willing to pay s_t only if

$$-(1-\delta)s_t + \delta\bar{\Pi} \geq \delta\underline{\Pi}. \tag{1}$$

For small $\epsilon > 0$, consider the following deviation by agent t. He chooses zero effort and the threat:

$$\mu_t(s) = \begin{cases} C & s = s_t - \epsilon \\ \\ D & \text{otherwise.} \end{cases} \tag{2}$$

Since (1) holds weakly at s_t, it holds strictly for $s_t - \epsilon$ and so the principal's unique best response to this deviation is to pay $s_t - \epsilon$. Agent t's payoff from this deviation is therefore $s_t - \epsilon$, which is strictly larger than $s_t - c(e_t)$ for sufficiently small ϵ. Hence, agent t can profitably deviate from any $e_t > 0$. Every equilibrium therefore has $e_t = 0$ for all $t \geq 0$, in which case $\bar{\Pi} = \underline{\Pi} = 0$ and so $s_t = 0$. ■

Whenever $s_t > 0$ on the equilibrium path, agent t can shirk and threaten to send a message that punishes the principal unless she pays him *slightly less* than s_t. Since the principal is willing to pay s_t to avoid this punishment, she strictly prefers to pay a smaller amount. Agent t can therefore shirk and still guarantee nearly the same transfer as if he had exerted effort. This deviation is so tempting that no agent will work.

Before moving on, we reflect on what Proposition 2 reveals about the economics of extortion. Any equilibrium specifies a mapping from agent t's messages to the principal's continuation payoffs. Let $\bar{\Pi}$ and $\underline{\Pi}$ be, respectively, the largest and smallest continuation payoffs in the image of this mapping. Agent t's gain from extortion depends on his **leverage** over the principal, defined as the normalized difference between these continuation payoffs,

$$L \equiv \frac{\delta}{1-\delta}\left(\bar{\Pi} - \underline{\Pi}\right). \tag{3}$$

In the no-extortion game, the principal pays $s_t = 0$ following any deviation and pays some $s_t \leq L$ on the equilibrium path. Increasing agent t's leverage therefore unambiguously increases the scope for cooperation. In the extortion game, on the other hand, $s_t \approx L$ both on- and off-path. Proposition 2 follows because agent t's leverage L, and so the transfer that he can demand, is independent of his effort.

This argument suggests that agent t *would* have the incentive to exert effort if doing so would increase his leverage and hence the pay that he could demand. The rest of the paper explores this idea: to deter extortion, tie leverage to effort. As the next sections demonstrate, tying leverage to effort requires that we construct coordinated punishments in a fundamentally different way.

4 Investigations

This section considers public signals of efforts or transfers. In the no-extortion game, such signals would be irrelevant; transfer signals would be redundant with the agents' messages, while effort signals would be redundant with what the principal, who is the only player that can directly punish shirking, already observes.[6]

However, these signals do have the potential to deter extortion. We first show that effort signals can tie an agent's leverage to his effort in equilibrium, which can induce effort. However, deterring extortion in this way requires agents to earn rent, creating a tension between the surplus created in equilibrium and the surplus captured by the principal. Then, we show that transfer signals usually cannot tie leverage to effort. Therefore, transfer signals

[6] Formally, the effort level in Proposition 1 is the highest attainable effort even if we drop all equilibrium constraints except for the principal's self-enforcement constraint and the agents' participation constraints. Those two sets of constraints would be unaffected by signals.

improve cooperation only under stringent conditions.

In the context of the GM-Fremont plant discussed in the introduction, our analysis suggests that unions should investigate the worker who files a grievance (i.e., effort), rather than just the subject of the grievance itself (i.e., the transfer). We will show that a grievance should optimally trigger harsher punishments when the investigation reveals that the filing worker has exerted more effort. Such an investigatory process can improve productivity, leading to better outcomes for both workers and the manager.[7] Similarly, we should observe workers exerting more effort in settings where doing so improves their ability to trigger coordinated punishments.

4.1 Effort Investigations

The **extortion game with effort signals** is similar to the baseline extortion game, except that an effort-dependent signal, y_t, is publicly observed after s_t. We focus on a simple, binary signal structure: $y_t \in \{0, 1\}$ with $\Pr\{y_t = 1 | e_t\} = \gamma(e_t)$ for $\gamma(\cdot)$ strictly increasing and twice continuously differentiable. Agent t's threat can be any mapping from his pay *and* this signal to a message, so that (with an abuse of notation) $\mu_t : \mathbb{R}^2 \to M$ and $m_t = \mu_t(s_t, y_t)$. Payoffs are the same as in the extortion game.

The signal y_t can deter extortion by making an agent's expected leverage an increasing function of his effort. Because signals are noisy, however, a shirking agent typically retains some leverage and, hence, can extort some pay. Agents therefore refrain from extortion only if they earn an equilibrium rent. Our result for this section shows how the tension between total surplus and the agents' rents determines effort in a principal-optimal equilibrium.

Proposition 3 *Consider an equilibrium of the game with effort signals. If $e_t = e$ on the*

[7] A practical caveat: this investigation must be made immune to manipulation by the manager, since she has the incentive to fabricate evidence of shirking in order to ensure that the grievance is ignored.

equilibrium path, then agent t's equilibrium payoff is at least $\bar{u}(e)$, where

$$\bar{u}(e) \equiv \max\left\{0, \frac{c'(e)}{\gamma'(e)}\gamma(e) - c(e)\right\}.$$

Suppose $\gamma(\cdot)$ is weakly concave. Then, $\bar{u}(\cdot)$ is strictly increasing, and in any $t \geq 0$ of any principal-optimal equilibrium, on-path effort solves

$$e_t \in \arg\max_e \{e - c(e) - \bar{u}(e)\}$$

subject to the constraint

$$\frac{c'(e)}{\gamma'(e)} \leq \frac{\delta}{1-\delta}(e - c(e) - \bar{u}(e)). \tag{4}$$

Proof: See Appendix A.

To prove Proposition 3, let $\bar{\Pi}(y)$ and $\underline{\Pi}(y)$ be the largest and smallest continuation payoffs induced by some message when the signal equals y. We can define an agent's leverage, $L(y)$, analogously to (3). Then, expected leverage, $\mathbb{E}\left[L(y)|e\right]$, depends on effort. As in Proposition 2, agent t can extort any transfer that is smaller than his expected leverage, so he chooses e_t to solve

$$e_t \in \arg\max_e \left\{\mathbb{E}\left[L(y)|e\right] - c(e)\right\}. \tag{5}$$

Since $L(\cdot) \geq 0$, this incentive constraint is identical to that of a static moral-hazard problem with limited liability; agent t's leverage is the analogue of the contractual payment, which can depend on y. As is typical in such models, agent t earns a rent, which equals $\bar{u}(e_t)$ for this signal structure.

As in a static moral-hazard problem with limited liability, it is optimal to set $L(0) = 0$; that is, agent t's message affects the principal's continuation payoff only if $y_t = 1$. If $\gamma(\cdot)$ is concave, then we can replace (5) with its first-order condition, $L(1) = c'(e)/\gamma'(e)$. Calculating the principal-optimal equilibrium payoff therefore reduces to maximizing total surplus minus the agent's rent, given that $L(1)$ cannot exceed the principal's equilibrium continuation payoff. Moreover, the principal's on-path continuation payoff equals her maximum equilibrium payoff, since we could otherwise increase it without affecting $L(\cdot)$. Thus, $L(1) = c'(e)/\gamma'(e)$ must satisfy the dynamic enforcement constraint, (4).

One immediate consequence of Proposition 3 is that there exists a principal-optimal equilibrium that is stationary on the equilibrium path. A second consequence is that agent t's maximum leverage is limited by the fact that *future* agents earn rent in equilibrium. That is, the right-hand side of (4) is decreasing in $\bar{u}(\cdot)$, which implies that each agent's rent-seeking behavior imposes a negative externality on the principal's relationships with other agents.

In practice, agents might have some sway over the signal distribution, as, for instance, when a union decides how to investigate grievances. Both the principal and agents prefer some kind of investigation to none, but they disagree on the optimal signal structure. In a principal-optimal equilibrium, the principal's payoff is maximized by the signal distribution that maximizes $e - c(e) - \bar{u}(e)$, while the agent's payoff is maximized by the distribution that maximizes $\bar{u}(e)$. For fixed e, $\bar{u}(e)$ is increasing in $\frac{\gamma(e)}{\gamma'(e)}$, which is larger when the signal distribution puts weight on "false positives:" $y_t = 1$ occurs frequently and with a probability that is (locally) not very responsive to effort. Thus, agents might collectively benefit from investigations that occasionally generate false positives, though of course, excessive false positives can undermine effort and lead to lower equilibrium rent.

4.2 Transfer Investigations

We now turn to public signals of transfers. In contrast to section 4.1, transfer signals are not a reliable remedy to extortion. The reason is that such signals reveal nothing about effort, so they usually cannot tie leverage to effort. The only exception is that certain signal distributions can be used to make the principal exactly indifferent between two different transfers when faced with an agent's optimal threat. This indifference allows us to construct equilibria in which agents have less leverage if they shirk.

The **extortion game with transfer signals** is identical to the extortion game except that in each period $t \geq 0$, a public signal $x_t \in \mathbb{R}$ is realized after s_t and observed by everyone. Agent t's threat maps each (s_t, x_t) to a message m_t, so $\mu_t : \mathbb{R}^2 \to M$ with $\mu_t(s_t, x_t) = m_t$. We again focus on binary signals, so that $x_t \in \{0, 1\}$ with $\Pr\{x_t = 1|s_t\} = \phi(s_t)$ for some strictly increasing and twice continuously differentiable $\phi(\cdot)$.

Our main result in this section is a set of necessary conditions on $\phi(\cdot)$ that must hold for an equilibrium with positive effort to exist. To understand these conditions, consider play in some period t. Define $\Pi(m_t, x_t)$ as the principal's continuation payoff if agent t's message is m_t and the signal is x_t. After agent t chooses his threat μ_t, the principal chooses s_t to maximize her payoff:

$$\max_s -(1-\delta)s + \delta\mathbb{E}\left[\Pi(\mu_t(s,x),x)|s\right]. \tag{6}$$

Note that (6) is independent of agent t's effort. Therefore, if a unique transfer maximizes (6), then the principal will pay that transfer regardless of agent t's effort. This leads to our first necessary condition: agent t exerts positive effort only if the principal is exactly indifferent between at least two transfers when she faces the equilibrium threat. The second necessary condition requires that no alternative threat would induce the principal to pay agent t more than his equilibrium payoff. Only under these two conditions is agent t willing to exert effort, and even then, the effort cost cannot exceed the difference between the on-path transfer and the largest amount that a shirking agent t can extort.

These two requirements imply a set of stringent necessary conditions on $\phi(\cdot)$.

Proposition 4 *Consider an equilibrium of the game with transfer signals. If $e_t > 0$ on the equilibrium path, then there exists $s^* > 0$ and $\hat{s} \in [0, s^*)$ such that (i) $c(e_t) \leqslant s^* - \hat{s}$, (ii) $\phi''(s^*) \leqslant 0$, and (iii)*

$$\phi'(s^*) = \frac{\phi(s^*) - \phi(\hat{s})}{s^* - \hat{s}}. \tag{7}$$

In particular, if $\phi(\cdot)$ is strictly concave on $\mathbb{R}_+$, then $e_t = 0$ in each $t \geq 0$ of every equilibrium.

Equation (7) combines the two conditions for $e_t > 0$ described above. First, the principal must be indifferent between paying the on-path transfer, s^*, and some other amount that is no less than $\hat{s}$, when faced with the equilibrium threat. Second, *no* threat can induce the principal to pay a transfer near s^*. The first of these conditions pins down the average slope of $\phi(\cdot)$ between $\hat{s}$ and s^*, while the second condition says that the derivative of $\phi(\cdot)$ near s^* equals the same number. Therefore, the average slope between $\hat{s}$ and s^* must equal the tangent slope at s^*, implying (7). Period-t effort must then satisfy $s^* - c(e_t) \geq \hat{s}$, since otherwise agent t could profitably shirk and extort $\hat{s}$.

Condition (7) cannot hold if $\phi(\cdot)$ is strictly concave, in which case every equilibrium entails $e_t = s_t = 0$ in each $t \geq 0$, just as in the extortion game without transfer signals. Thus, positive equilibrium effort is possible only if $\phi(\cdot)$ has both convex and concave regions. For particular examples of signal structures, we can construct equilibria with positive effort. Such equilibria require the principal to be punished whenever $x_t = 0$, since otherwise $\mathbb{E}\left[\Pi(\mu_t(s_t, x_t), x_t)|s_t\right]$ would be constant in s_t. The principal is therefore periodically punished on the equilibrium path in any equilibrium with positive effort. For these reasons, we view transfer investigations as unreliable, in the sense that they do not improve cooperation for a wide variety of signal distributions, and inefficient, because any equilibrium with positive effort must also entail occasional on-path punishments.

5 Dyadic Relationships

In the extortion game, the principal can punish an agent only by withholding pay, while an agent can punish the principal only by communicating with future agents. In this section, we explore how ongoing interactions between the principal and each individual agent can deter extortion. As is familiar from the literature on repeated games, these *dyadic* relationships can be used to punish an agent for shirking or the principal for reneging on a hard-working agent. We now emphasize a third effect that is new to our setting: dyadic relationships can be used to punish the principal for acquiescing to extortion, which decreases the leverage of a shirking agent. By tying leverage to effort in this way, dyadic relationships enable coordinated punishments.

Consider the **extortion game with dyadic relationships**, which makes two changes to the extortion game. The first is minor: when the principal chooses s_t, we allow agent t to simultaneously make a transfer to the principal, $s_t^A \geq 0$, which is observed by the principal but not by other agents. Note that allowing such transfers would not change any of our other results. The second, more substantial change is that *after* agent t sends his message in each period $t \geq 0$, the principal and agent t play a symmetric, simultaneous-move coordination game. The actions and outcomes of this coordination game are observed by the two participants but not by any other agents. We suppress notation for actions in this coordination game and instead denote the resulting (symmetric) payoff by v_t, so that the principal's and agent t's payoffs are $e_t - s_t + s_t^A + v_t$ and $s_t - s_t^A - c(e_t) + v_t$, respectively.

The outcome of each coordination game is not observed by future agents and so cannot affect the principal's continuation payoff. In equilibrium, v_t must therefore correspond to a Nash equilibrium of the coordination game in each $t \geq 0$. Define $v_t = v_H$ and $v_t = v_L$ as the largest and smallest such Nash equilibrium payoffs, respectively. While our result can be readily extended for general, asymmetric coordination games, the following simple game suffices:

$$\begin{array}{ccc} & h & l \\ h & (v_H, v_H) & (v_L, v_L) \\ l & (v_L, v_L) & (v_L, v_L) \end{array} \cdot$$

We show that positive effort can be sustained in the extortion game with dyadic relationships. However, equilibrium effort is constrained by the strength of each dyadic relationship, as measured by the difference $(v_H - v_L)$.

Proposition 5 *In the extortion game with dyadic relationships, $c(e_t) \leq 3(v_H - v_L)$ in every $t \geq 0$ of any equilibrium. If e^* is the minimum of e^{FB} and the solution to $c(e^*) = 3(v_H - v_L)$, then there exists a $\bar{\delta} < 1$ such that for any $\delta \geq \bar{\delta}$, $e_t = e^*$ in every $t \geq 0$ on the equilibrium path in any principal-optimal equilibrium.*

Proof: See appendix A.

The constraint $c(e_t) \leq 3(v_H - v_L)$ reflects the fact that dyadic relationships optimally encourage cooperation via three channels: they (i) punish agents for shirking, (ii) punish the principal for refusing to pay a hard-working agent, and (iii) *reward* the principal for refusing to pay a shirking agent. The first two of these channels are familiar. The third channel is new and shows how dyadic relationships enable coordinated punishments.

Adopting the notation from Section 3, an agent's on-path leverage equals

$$L^* = \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}) + (v_H - v_L),$$

reflecting the fact that a principal who reneges is punished in both the period-t coordination game and the continuation equilibrium. If agent t shirks, then his leverage decreases to

$$\hat{L} = \max\left\{\frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}) - (v_H - v_L), 0\right\},$$

since the coordination game rewards the principal for not paying a shirking agent. An agent can extort any transfer that is strictly less than his leverage, so his on-path transfer is $L^* - \hat{L}$ larger than the maximum amount he can extort. This difference is maximized if

$$\frac{\delta}{1-\delta}\left(\bar{\Pi} - \underline{\Pi}\right) = v_H - v_L, \tag{8}$$

in which case $L^* - \hat{L} = 2(v_H - v_L)$. Combining this difference in transfers with the fact that a shirking agent faces a direct punishment of $(v_H - v_L)$, we conclude that equilibrium effort must satisfy $c(e^*) \leq 3(v_H - v_L)$.

As is familiar from the literature on cooperation, dyadic relationships can encourage cooperation by increasing on-path leverage, L^*, and by directly punishing a shirking agent. Papers that focus on how dyadic relationships might be misused, including Basu (2003), Dixit (2003a), and Myerson (2004), also consider these two channels. In contrast, we focus on the third channel, which is new to our setting: dyadic relationships can complement coordinated punishments by decreasing the leverage of a shirking agent, $\hat{L}$.

Decreasing $\hat{L}$ would be irrelevant in a setting without extortion, since in that case, the value of coordinated punishments depends only on how they affect on-path leverage, L^*. In the extortion game, however, what matters is how leverage varies with effort, $L^* - \hat{L}$. Decreasing $\hat{L}$ means that each agent can be given access to coordinated punishments without misusing them. Consequently, as represented by (8), stronger dyadic relationships (measured by $v_H - v_L$) optimally expand the scope for coordinated punishments (measured by $\bar{\Pi} - \underline{\Pi}$).

The coordination game is an abstract way to capture the idea that the principal has ongoing interactions with each agent. In reality, managers interact repeatedly with each of their employees, community members have repeated opportunities to contribute to public goods, and most businesses are long-term members of their associations. We interpret the coordination-game payoff, v_t, as a simple representation of the continuation payoff from these future interactions. Our result crystallizes the idea that organizations with stronger dyadic relationships can better deter extortion. In Appendix D, we confirm this interpretation by studying a setting with long-run agents who interact repeatedly with the principal. While the resulting analysis is more involved, it remains true that dyadic relationships facilitate coordinated punishments.

Stepping back from the formal analysis, how might a firm cultivate dyadic relationships that deter extortion? The first step is to create manager-worker relationships with multiple equilibrium payoffs, so that $v_H - v_L$ is large. The firm's formal contracts must be structured in a way that supports this multiplicity (see, e.g., Che and Yoo (2001)). This was not the case at GM-Fremont, where managerial incentives were based heavily on formal contracts that left little room for relational contracts (Glass and Langfitt (2015)).[8] The organization's culture must then select among these equilibria in a way that deters extortion. By implementing the right formal incentives and fostering the right culture, an organization can encourage strong

[8] Both managers and workers were incentivized to keep the production line running at all times, so they had little incentive to cooperate on, e.g., fixing production mistakes or otherwise improving quality.

dyadic relationships that support effective coordinated punishments.

6 Spillovers from Extorting to Non-Extorting Agents

This section enriches our baseline model so that some, but not all, agents extort in principal-optimal equilibria. We show that extortion spills over onto non-extorting relationships in two ways. First, the possibility of future extortion makes the principal less willing to pay transfers today. Second, to make extortion less attractive, principal-optimal equilibria might implement weak coordinated punishments that lead to low effort from non-extorting agents.

Consider the following **costly extortion game.** Suppose that, at the start of every period $t \in \{0, 1, ...\}$, agent t privately observes a cost $k_t \geq 0$, $k_t \sim G(\cdot)$, and then chooses whether or not to invest. If he invests, then his payoff decreases by k_t and he plays the extortion game with the principal; otherwise, he plays the no-extortion game with the principal. Only the principal observes agent t's investment decision; other agents observe only m_t.

We interpret k_t as agent t's cost of making the principal believe that he will follow through on his threat. An agent might incur this cost by developing an (unmodeled) reputation for following through on extortionary threats or otherwise demonstrating that he is willing to extort. The extortion and the no-extortion games are special cases of this game where $k_t = 0$ or k_t is large, respectively. In this section, we focus on distributions over k_t such that agents invest with an interior probability.

We characterize principal-optimal equilibria in the costly extortion game. This result is phrased in terms of an agent's leverage in order to discuss how changing leverage changes the prevalence of extortion and its consequences for cooperation.

Proposition 6 *Consider the costly extortion game. At any on-path, period-t history of any principal-optimal equilibrium, there exists an L_t such that, if agent t invests, then $e_t = 0$ and*

$s_t = L_t$, while if he does not invest, then $e_t = c^{-1}(L_t)$ and $s_t = L_t$. Moreover, L_t solves

$$L_t \in \arg\max_L \left\{(1 - G(L))\, c^{-1}(L) - L\right\}$$

subject to the constraint

$$L \leq \frac{\delta}{1-\delta}\left((1 - G(L))\, c^{-1}(L) - L\right). \tag{9}$$

Proof: See Appendix A.

Using the notation from Section 3, define

$$L \equiv \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi})$$

as agent t's leverage. If agent t invests, then as in the proof of Proposition 2, his unique equilibrium strategy is to shirk and extort as much as possible, so the transfer equals L and effort equals zero. If agent t does not invest, then as in the proof of Proposition 1, he is willing to choose e_t only if $c(e_t) \leq s_t$, while the principal is willing to pay s_t only if $s_t \leq L$. Agent t invests whenever the costs of doing so, k_t, are smaller than the gains, $L-(s_t - c(e_t))$.

As in Proposition 3, principal-optimal equilibria are sequentially principal-optimal. In each period of such an equilibrium, L ensures that the principal is exactly willing to compensate each agent for his effort, given that (i) an agent who invests exerts zero effort and is paid L, and (ii) L is no more than the principal's equilibrium continuation payoff. These two conditions lead to (9).

Increasing an agent's leverage increases both his temptation to invest and the effort he is willing to exert if he does not. In a principal-optimal equilibrium, agent t extorts with probability $G(L)$ and otherwise exerts effort $e_t = c^{-1}(L)$. Thus, higher L has opposing effects on equilibrium cooperation: it leads to a higher prevalence of extortion and higher payments to extorting agents, but it also leads to higher effort among those agents who do not extort. The optimal L balances these forces and therefore limits the leverage available to non-extorting agents.

The dynamic enforcement constraint, (9), illustrates a further negative spillover from extorting to non-extorting relationships. The right-hand side of this constraint equals the principal's on-path continuation payoff. Future non-extorting agents contribute $c^{-1}(L)-L > 0$ to this payoff, while future extorting agents contribute $-L < 0$. Thus, even if extortion is not present in a given principal-agent relationship, and even if this fact is known to both parties, the possibility that *other* agents might extort undermines cooperation.

7 Conclusion

This paper studies a prevalent obstacle to using coordinated punishments to facilitate cooperation: agents may misuse messages intended to report deviations to extort the principal. We show that extortion has the potential to destroy cooperation. We also explore practical ways to restore cooperation, all of which build on the same core intuition: to deter misuse, tie an agent's leverage over the principal to his effort.

Communication is particularly susceptible to these kinds of extortionary threats for two reasons. First, communication is necessary precisely when players do not observe one another's interactions, which means that extortion is unlikely to be widely observed. Second, coordinated punishments are valuable when individual relationships are relatively weak, which means that the extorted party has little direct recourse to punish the extorter. Both of these features suggest that agents incur little cost from making, and following through on, extortionary threats.

Our analysis suggests three natural next steps. First, we could delve further into the assumption that agents commit to their threats. We have argued that this assumption is both reasonable and (in a sense) necessary to study extortion, which requires shirking agents to make action-contingent threats. We could therefore ask: how might an agent develop a reputation for following through on these threats? If such a reputation is publicly known, then other agents can simply ignore the resulting messages. If the principal is aware of the reputation and other agents are not, however, then a reputation for extortion is valuable. Each agent therefore has the incentive to develop a *public* reputation for honesty and a *bilateral* reputation for extortion.

Second, we could consider how an organization's use of coordinated punishments affects the workers that it attracts and retains. An agent who is willing to extort the principal benefits more from coordinated punishments than agents who use those punishments only for their intended purpose. Therefore, organizations that rely on coordinated punishments risk attracting exactly those agents who are most likely to misuse them. How might an organization that relies on coordinated punishments overcome this adverse selection problem?

Finally, extortionary threats are also a feature in more symmetric interactions, as in, for example, communal enforcement (e.g., Dixit (2007), Ali and Miller (2016)). In such settings, *both* sides have the opportunity to extort one another. How do players cooperate in the presence of two-sided extortion? What networks best facilitate cooperation, and how are rents shared within those networks? How should business associations, communities, and firms structure their communication channels to support strong relational contracts? We hope that our analysis provides a foundation for analyzing such questions.

References

Ali, S. N. and C. Liu (2018). Conventions and coalitions in repeated games. Working Paper.

Ali, S. N. and D. Miller (2013). Enforcing cooperation in networked societies. Working Paper.

Ali, S. N. and D. Miller (2016). Ostracism and forgiveness. *American Economic Review 106*(8), 2329–2348.

Ali, S. N., D. Miller, and D. Yang (2017). Renegotiation-proof multilateral enforcement.

Andrews, I. and D. Barron (2016). The allocation of future business: Dynamic relational contracts with multiple agents. *American Economic Review 106*(9), 2742–2759.

Arnold, C. and R. Smith (2016, 10). Bad form, wells fargo. NPR.

Baker, G., R. Gibbons, and K. Murphy (1994). Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics 109*(4), 1125–1156.

Baker, G., R. Gibbons, and K. J. Murphy (2002). Relational contracts and the theory of the firm. *The Quarterly Journal of Economics 117*(1), 39–84.

Barron, D., J. Li, and M. Zator (2018). Productivity and debt in relational contracts. Working Paper.

Barron, D. and M. Powell (2018). Policies in relational contracts. Forthcoming, American Economic Journal: Microeconomics.

Basu, K. (2003). *Analytical Development Economics: the Less Developed Economy Revisited*. MIT Press.

Bernstein, L. (2015). Beyond relational contracts: Social capital and network governance in procurement contracts. *Journal of Legal Analysis 7*(2), 561–621.

Board, S. (2011). Relational contracts and the value of loyalty. *American Economic Review 101*(7), 3349–3367.

Bowen, T. R., D. M. Kreps, and A. Skrzypacz (2013). Rules with discretion and local information. *The Quarterly Journal of Economics 128*(3), 1273–1320.

Bull, C. (1987). The existence of self-enforcing implicit contracts. *The Quarterly Journal of Economics 102*(1), 147–159.

Chassang, S. and G. Padro i Miquel (2018). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. Forthcoming, Review of Economic Studies.

Che, Y.-K. and S.-W. Yoo (2001). Optimal incentives for teams. *The American Economic Review 91*(3), 525–541.

Conti, A. (2019). I accidentally uncovered a nationwide scam on airbnb. *Vice*.

Dewatripont, M. (1987). The role of indifference in sequential models of spatial competition: an example. *Economics Letters 23*(4), 323–328.

Dixit, A. (2003a). On modes of economic governance. *Econometrica 71*(2), 449–481.

Dixit, A. (2003b). Trade expansion and contract enforcement. *Journal of Political Economy 111*(6), 1293–1317.

Dixit, A. (2007). *Lawlessness and Economics: Alternative Modes of Governance*. Princeton University Press.

Fong, Y.-F. and J. Li (2017). Relational contracts, limited liability, and employment dynamics.

Freeman, R. and J. Medoff (1979). The two faces of unionism. *The Public Interest 57*, 69–93.

Gambetta, D. (1993). *The Sicilian Mafia: The Business of Private Protection*. Harvard University Press.

Gibbons, R. and R. Henderson (2013). What do managers do? exploring persistent performance differences among seemingly similar enterprises. In R. Gibbons and J. Roberts (Eds.), *Handbook of Organizational Economics*, pp. 680–731.

Glass, I. and F. Langfitt (2015). Nummi 2015.

Goldlucke, S. and S. Kranz (2017). Reconciliating relational contracting and hold-up: A model of repeated negotiations. Working Paper.

Greif, A., P. Milgrom, and B. Weingast (1994). Coordination, commitment, and enforcement: The case of the merchant guild. *Journal of Political Economy 102*(4), 745–776.

Guo, Y. and J. Hörner (2018). Dynamic allocation without money. Working Paper.

Halac, M. (2012). Relational contracts and the value of relationships. *American Economic Review 102*(2), 750–779.

Halac, M. (2015). Investing in a relationship. *RAND Journal of Economics 46*(1), 165–186.

Hörner, J. and N. Lambert (2018). Motivational ratings. *Review of Economic Studies*. Forthcoming.

Klein, T., C. Lambertz, and K. Stahl (2016). Market transparency, adverse selection, and moral hazard. *Journal of Political Economy 124*(6), 1677–1713.

Levin, J. (2002). Multilateral contracting and the employment relationship. *The Quarterly Journal of Economics 117*(3), 1075–1103.

Levin, J. (2003). Relational incentive contracts. *The American Economic Review 93*(3), 835–857.

Li, J., N. Matouschek, and M. Powell (2017, February). Power dynamics in organizations. *American Economic Journal: Microeconomics 9*(1), 217–41.

Lipnowski, E. and J. Ramos (2017). Repeated delegation.

Lippert, S. and G. Spagnolo (2011). Networks of relations and word-of-mouth communication. *Games and Economic Behavior 72*(1), 202–217.

Liu, C. (2019). Stability in repeated matching markets. Working Paper.

MacLeod, B. and J. Malcomson (1989). Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica 57*(2), 447–480.

Malcomson, J. (2013). Relational incentive contracts. In R. Gibbons and J. Roberts (Eds.), *Handbook of Organizational Economics*, pp. 1014–1065.

Malcomson, J. (2016). Relational contracts with private information. *Econometrica 84*(1), 317–346.

Milgrom, P., D. North, and B. Weingast (1990). The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics and Politics 2*(1), 1–23.

Miller, D., T. Olsen, and J. Watson (2018). Relational contracting, negotiation, and external enforcement. Working Paper.

Miller, D. and J. Watson (2013). A theory of disagreement in repeated games with bargaining. *Econometrica 81*(6), 2303–2350.

Myerson, R. B. (2004). Justice, institutions, and multiple equilibria. *Chicago Journal of International Law 5*(1), 91–108.

Ortner, J. and S. Chassang (2018). Making corruption harder: Asymmetric information, collusion, and crime. *Journal of Political Economy 126*(5), 2108–2133.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

Tranaes, T. (1998). Tie-breaking in games of perfect information. *Games and Economic Behavior 22*(1), 148–161.

Watson, J. (2017). A general, practicable definition of perfect bayesian equilibrium.

Wolitzky, A. (2012). Career concerns and performance reporting in optimal incentive contracts. *B.E. Journal of Theoretical Economics (Contributions) 12*(1).

Wolitzky, A. (2013). Cooperation with network monitoring. *The Review of Economic Studies 80*(1), 395–427.

Zhu, J. Y. (2018). A foundation for efficiency wage contracts. *American Economic Journal: Microeconomics 10*(4), 248–288.

Zhu, J. Y. (2019). Better monitoring...worse productivity? Working Paper.

A Omitted Proofs

A.1 Proof of Proposition 3

Consider an equilibrium. Suppose $e_t = e$ at some on-path, period-t history, and let $\bar{\Pi}(y)$ and $\underline{\Pi}(y)$ be the principal's largest and smallest continuation payoffs following signal realization y, with corresponding messages $\bar{m}(y)$ and $\underline{m}(y)$. Define $L(y) \equiv \frac{\delta}{1-\delta}(\bar{\Pi}(y) - \underline{\Pi}(y))$.

For each effort e_t, agent t can choose

$$\mu_t(s, y) = \begin{cases} \bar{m}(y) & s_t \geq \hat{s} \\ \underline{m}(y) & \text{otherwise.} \end{cases}$$

Whenever

$$\hat{s} < \hat{s}(e_t) \equiv L(0) + \gamma(e_t)(L(1) - L(0)),$$

the principal's unique best response to this μ_t is to pay $\hat{s}$. On the other hand $s_t = 0$ is a best response to any $\hat{s} \geq \hat{s}(e_t)$. Thus, agent t's equilibrium effort, e, must satisfy

$$e \in \arg\max_{e'} \left\{ \hat{s}(e') - c(e') \right\}.$$

If $e > 0$, then a necessary condition for agent t to choose $e_t = e$ is that

$$c'(e) = \hat{s}'(e) = \gamma'(e)(L(1) - L(0)). \tag{10}$$

Since $\gamma'(e) > 0$, we can solve for $L(1) - L(0)$ in (10) and plug into the definition of $\hat{s}(e_t)$ to yield

$$\hat{s}(e) \geq L(0) + \gamma(e)\frac{c'(e)}{\gamma'(e)}.$$

Agent t earns at least 0, so

$$s_t - c(e) \geq \max\{0, \hat{s}(e) - c(e)\} = \max\left\{0, \gamma(e)\frac{c'(e)}{\gamma'(e)} - c(e)\right\} \equiv \bar{u}(e),$$

as desired.

Now, suppose $\gamma(\cdot)$ is concave. Since $c'(0) = c(0) = 0$, $\bar{u}(0) = 0$, and

$$\frac{d}{de}\left\{\gamma(e)\frac{c'(e)}{\gamma'(e)} - c(e)\right\} > 0,$$

so that $\bar{u}(\cdot)$ is strictly increasing. Moreover, the first-order condition (10) is both necessary and sufficient for agent t to exert effort $e_t = e$.

We now characterize principal-optimal equilibrium. Let Π^* be the principal's payoff in such an equilibrium. Note that on the equilibrium path, the principal's continuation payoff equals $\bar{\Pi}(y)$ following realization y, since otherwise agent t could demand a higher transfer using the promise of $\bar{\Pi}(y)$.

Suppose that $\bar{\Pi}(y) < \Pi^*$ for some $y \in \{0, 1\}$. In that case, we can increase both $\bar{\Pi}(y)$ and $\underline{\Pi}(y)$ by the same constant to keep $L(y)$, and hence agent t's incentives, unchanged. Doing so strictly increases the principal's payoff. So the principal's on-path continuation payoff equals Π^* in each $t \geq 0$ of any principal-optimal equilibrium. But then $\Pi^* = (1-\delta)(e_t - s_t) + \delta\Pi^*$, so $\Pi^* = e_t - s_t$ in any $t \geq 0$ on the equilibrium path.

In a principal-optimal equilibrium with $\gamma''(e) \leq 0$, $s_t = \mathbb{E}[L(y)|e]$, where e_t solves

$$\max_{L(\cdot)\geq 0, e} e - \mathbb{E}[L(y)|e]$$

subject to (10) and

$$L(y) \leq \frac{\delta}{1-\delta}\Pi^*.$$

Thus, $L(0) = 0$, in which case $L(1) = \frac{c'(e)}{\gamma'(e)}$ and so $\mathbb{E}\left[L(y)|e]\right] = \gamma(e)\frac{c'(e)}{\gamma'(e)} = \bar{u}(e) + c(e)$. Moreover, $\Pi^* = e_t - s_t$ by the argument above, where e_t and s_t solve an identical constrained maximization problem. Substituting these simplifications into this constrained maximization problem yields the constrained maximization problem in the statement of the Proposition. ■

A.2 Proof of Proposition 4

Fix a period t. Let $\Pi(m, x)$ be the principal's continuation payoff following message m and signal x. Let $\bar{\Pi}(x) = \max_m \Pi(m, x)$ and $\underline{\Pi}(x) = \min_m \Pi(m, x)$ with $\bar{m}(x)$ and $\underline{m}(x)$ being the corresponding maximizer and minimizer. We let π^D be the smallest payoff that the principal can guarantee herself,

$$\pi^D = \max_s -(1-\delta)s + \delta\mathbb{E}\left[\underline{\Pi}(x)|s\right]. \tag{11}$$

Define s_A as the smallest maximizer of (11). We argue that agent t's payoff is at least s_A. He can always choose $e_t = 0$ and

$$\mu_t(s, x) = \begin{cases} \bar{m}(x), & \text{if } s = s_A \\ \underline{m}(x), & \text{if } s \neq s_A. \end{cases}$$

Faced with this threat, the principal earns π^D from paying s_A and strictly less than π^D from paying $s < s_A$. Therefore, the principal will pay at least s_A.

Consider the set of transfers that can give the principal a higher payoff than π^D:

$$\{s : -(1-\delta)s + \delta\mathbb{E}\left[\bar{\Pi}(x)|s\right] > \pi^D\}. \tag{12}$$

If this set is nonempty, we let s_B be the supremum of this set. We argue that agent t can get a payoff arbitrarily close to s_B. In particular, he can choose $e_t = 0$ and

$$\mu_t(s,x) = \begin{cases} \bar{m}(x), & \text{if } s = s_B - \epsilon \\ \underline{m}(x), & \text{if } s \neq s_B - \epsilon. \end{cases}$$

Since $\phi(\cdot)$ is continuous, the principal's unique best response is to pay $s_t = s_B - \epsilon$ for small enough $\epsilon > 0$.

Now, define $\hat{s} = \max\{s_A, s_B\}$ if the set (12) is nonempty, and $\hat{s} = s_A$ otherwise. Agent t can guarantee a payoff arbitrarily close to $\hat{s}$ if he shirks, so he chooses $e_t = e^*$ only if $s^* - c(e^*) \geqslant \hat{s}$, which is our first necessary condition. Moreover, we can show that

$$-(1-\delta)s^* + \delta\mathbb{E}\left[\bar{\Pi}(x)|s^*\right] = s^D \tag{13}$$

$$-(1-\delta)\hat{s} + \delta\mathbb{E}\left[\bar{\Pi}(x)|\hat{s}\right] = s^D. \tag{14}$$

To see why (13) holds, note that the principal is willing to pay s^* so the left-hand side of (13) must be weakly higher than π^D. But either s^B does not exist, in which case (13) must hold with equality, or the supremum of the set (12) must be strictly below s^*, so that again (13) holds with equality. Equality (14) follows from the continuity of $\phi(\cdot)$ and the definition of $\hat{s}$.

Combining (13) and (14), we have

$$s^* - \hat{s} = \frac{\delta}{1-\delta}\left(\phi(s^*) - \phi(\hat{s})\right)\left(\bar{\Pi}(1) - \bar{\Pi}(0)\right). \tag{15}$$

Given (13) , $-(1-\delta)s + \delta\mathbb{E}\left[\bar{\Pi}(x)|s\right]$ must attain a local maximum at $s = s^*$, since otherwise (12) would contain elements arbitrarily close to s^* and so $s^* \leq \hat{s}$. Thus,

$$\phi'(s^*)\left(\bar{\Pi}(1) - \bar{\Pi}(0)\right) = \frac{1-\delta}{\delta} \tag{16}$$

and $\phi''(s) \leqslant 0$. Combining (15) and (16) yields our final necessary condition:

$$\phi'(s^*) = \frac{\phi(s^*) - \phi(\hat{s})}{s^* - \hat{s}}.$$

If $\phi(\cdot)$ is strictly concave, it cannot satisfy this condition for $s^* > \hat{s}$. ■

A.3 Proof of Proposition 5

Consider period t of an equilibrium. Define $\bar{\Pi}$ and $\underline{\Pi}$ as the principal's largest and smallest continuation payoffs, respectively, with corresponding messages $\bar{m}$ and $\underline{m}$. Agent t can always deviate to $e_t = s_t^A = 0$ and

$$\mu_t(s) = \begin{cases} \bar{m} & s = \hat{s} \\ \underline{m} & \text{otherwise.} \end{cases}$$

Following this deviation, the principal's unique best response is $s_t = \hat{s}$ if

$$\hat{s} < v_L - v_H + \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}). \tag{17}$$

Similarly, if agent t does not deviate, the principal is willing to pay $s_t = s^*$ only if

$$s^* \leq v_H - v_L + \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}). \tag{18}$$

Agent t is willing to choose $e_t = e^*$ only if $s^* - c(e^*) + (v_H - v_L) \geq \hat{s}$ for *any* $\hat{s}$ satisfying (17). Given the bound (18) on s^*, we conclude that $e_t = e^*$ in equilibrium only if $3(v_H - v_L) \geq c(e^*)$.

Each agent must earn at least v_L, so the principal's equilibrium payoff cannot exceed $e^* - c(e^*) + 2v_H - v_L$, where $e^* = e^{FB}$ if $c(e^{FB}) \leq 3(v_H - v_L)$ and e^* satisfies $c(e^*) = 3(v_H - v_L)$ otherwise. To complete the proof, we construct an equilibrium that attains this bound. Play starts in the cooperative phase: in each $t \geq 0$, agent t chooses $e_t = e^*$ and

$$\mu_t(s) = \begin{cases} C & s = c(e^*) \\ D & \text{otherwise.} \end{cases}$$

Transfers equal $s_t = \min\{0, c(e^*) - (v_H - v_L)\}$, $s_t^A = \min\{0, (v_H - v_L) - c(e^*)\}$ if agent t does not deviate and $s_t = 0$, $s_t^A = (v_H - v_L)$ if he does. If either nobody deviates or agent t deviates from (e_t, μ_t) but then nobody deviates from s_t, then $v_t = v_H$; otherwise, $v_t = v_L$. Play continues in the cooperative phase until $m_t = D$, at which point it transitions to the punishment phase with probability α. In the punishment phase, $e_t = s_t = 0$ in each period. Let α satisfy

$$\max\{0, c(e^*) - 2(v_H - v_L)\} = \frac{\delta}{1-\delta}\alpha\left(e^* - c(e^*) + 2v_H - v_L\right).$$

For $\delta < 1$ sufficiently close to 1, $\alpha \in [0,1]$.

The principal earns $e^* - c(e^*) + v_H + (v_H - v_L)$ surplus in each period of the cooperative phase. If agent t deviates in (e_t, μ_t), then he earns v_L by paying $s_t^A = (v_H - v_L)$ and $-s_t^A + v_L$ from deviating, so he has no profitable deviation from s_t^A. Regardless of μ_t, the principal has no profitable deviation from $s_t = 0$ following a deviation in (e_t, μ_t) if

$$v_H - v_L \geq \frac{\delta}{1-\delta}\alpha(e^* - c(e^*) + 2v_H - v_L) = \max\{0, c(e^*) - 2(v_H - v_L)\},$$

which holds because $c(e^*) \leq 3(v_H - v_L)$. On the equilibrium path, if $c(e^*) - (v_H - v_L) \geq 0$, then the principal has no profitable deviation from s_t because

$$-c(e^*) + (v_H - v_L) + v_H + \frac{\delta}{1-\delta}(e^* - c(e^*) + 2v_H - v_L) \geq v_L + \frac{\delta}{1-\delta}(1-\alpha)(e^* - c(e^*) + 2v_H - v_L).$$

This is because, by definition of α,

$$-c(e^*) + 2(v_H - v_L) \geq \frac{\delta}{1-\delta}\alpha(e^* - c(e^*) + 2v_H - v_L).$$

If $c(e^*) - (v_H - v_L) < 0$, then agent t has no profitable deviation from s_t^A because $c(e^*) - (v_H - v_L) + v_H \geq v_L$.

Given these transfers, agent t earns v_L from choosing the equilibrium (e_t, μ_t) and no more than v_L from deviating. So this strategy profile is an equilibrium. It is principal-optimal because it attains the upper bound on the principal's equilibrium payoff. ■

A.4 Proof of Proposition 6

Consider an equilibrium and a history at the start of period t. Define $\bar{\Pi}$ and $\underline{\Pi}$ as in the proof of Proposition 2, with corresponding messages $\bar{m}$ and $\underline{m}$, and let

$$L_t \equiv \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}).$$

Suppose agent t invests. If $e_t > 0$ or $s_t < L_t$, then the deviation from the proof of Proposition 2 is profitable for $\epsilon > 0$ sufficiently small. Consequently, $e_t = 0$ and $s_t = L_t$ whenever agent t invests.

Suppose agent t does not invest. He must earn at least a payoff of zero in equilibrium, so $s_t - c(e_t) \geq 0$. The principal must be willing to pay s_t, so $s_t \leq L_t$. For any e_t and s_t that satisfy these two constraints, consider the following strategy profile:

1. Agent t chooses e_t.
2. The principal pays s_t if agent t has not deviated and pays nothing otherwise.
3. Agent t sends $\bar{m}$ if no deviation has occurred and $\underline{m}$ otherwise.

If agent t chooses e_t, the principal is willing to pay s_t because $s_t \leq L_t$. If agent t deviates, then $m_t = \underline{m}$ regardless of the principal's action, so she pays nothing. Agent t is willing to choose e_t because $s_t \geq c(e_t)$. Thus, neither player has a profitable deviation from this strategy profile. We conclude that any (e_t, s_t) with $c(e_t) \leqslant s_t \leqslant L_t$ can be implemented in an equilibrium, as desired.

Since an investing agent exerts no effort and obtains a pay of L_t, from now on we use e_t, s_t for the effort exerted by, and the pay received by, agent t who didn't invest. Given this continuation play, agent t is willing to invest if and only if

$$L_t - (s_t - c(e_t)) \geq k_t,$$

where $L_t - (s_t - c(e_t))$ and k_t represent the gain from, and cost of, investment, respectively.

Now, consider a principal-optimal equilibrium, and let Π^* equal the principal's maximum equilibrium payoff. We must have $\bar{\Pi} = \Pi^*$ in each period t, since agent t's incentive depends only on L_t so we can increase $\bar{\Pi}, \underline{\Pi}$ while keeping L_t fixed. The principal's payoff is

$$\max_{L_t, s_t, e_t} G(L_t - (s_t - c(e_t))) \{\delta \Pi^* - (1-\delta) L_t\} + (1 - G(L_t - (s_t - c(e_t)))) \{(1-\delta)(e_t - s_t) + \delta \Pi^*\} \tag{19}$$

subject to the constraint that $c(e_t) \leqslant s_t \leqslant L_t$. The constraint $s_t \leqslant L_t$ must bind, since the principal would like L_t to be as small as possible. Substituting $L_t = s_t$ into (19), the objective in (19) becomes:

$$G(c(e_t)) \{\delta \Pi^* - (1-\delta) s_t\} + (1 - G(c(e_t))) \{(1-\delta)(e_t - s_t) + \delta \Pi^*\}$$

The derivative of this objective with respect to s_t is $-1 + \delta$. Hence, it is optimal to choose $s_t = c(e_t)$. The objective in (19) becomes

$$\delta \Pi^* + (1-\delta) \left((1 - G(c(e_t)) e_t - c(e_t) \right).$$

Therefore, the optimal effort maximizes $(1 - G(c(e_t))e_t - c(e_t)$ and Π^* is given by this maximum:

$$\Pi^* = \max_{e_t} (1 - G(c(e_t))e_t - c(e_t).$$

The formula for Π^* is quite clear. The principal has to pay $c(e_t)$ to both an extorting agent and a nonextorting one. However, she only obtains e_t from the nonexorting agent, which occurs with probability $1 - G(c(e_t))$. ■

B Online Appendix: Interpreting Commitment

In this section, we interpret the commitment assumption at the heart of our analysis.

Without the threat or a similar modeling device, Proposition 1 shows that we can always construct equilibria in which agents do not follow through on extortionary threats. Commitment is a straightforward way to make sure that agents' threats are more than just cheap talk. Crucially, however, the threat does not force an agent to send an *ex post* suboptimal message. Indeed, our next result shows that commitment refines the set of equilibria in each game that we study.

Recall that the no-extortion game is identical to the extortion game, except that each agent t chooses m_t freely at the end of period t rather than being committed to μ_t.

Proposition 7 *For any equilibrium of the extortion game or of the extortion game with effort signals, transfer signals, or dyadic relationships, there exists an equilibrium of the corresponding no-extortion game that induces the same distribution over* $(e_t, s_t, m_t)_{t=0}^{\infty}$.

Proof: In the extortion game, this result follows immediately from the fact that agents are indifferent among messages and so are willing to follow their threats. Proposition 2 shows such an equilibrium exists, which completes the proof. In the games with effort signals or transfer signals, agents are again indifferent over messages and so a nearly identical argument proves the result.

Consider the extortion game with bilateral relationships. Let σ^* be an equilibrium, and consider the following strategy profile of the game: in each period $t \geqslant 0$,

1. Agent t chooses e_t, μ_t as in σ^*.

2. The principal chooses s_t as in σ^*.

3. Agent t chooses $m_t = \mu_t(s_t)$.

4. If agent t follows this message strategy, a_t is as in σ^*; otherwise, $a_t = L$.

No player has a profitable deviation from a_t because a_t is always an equilibrium of the simultaneous move game at the end of the period. By the choice of a_t following a deviation in m_t, agent t has a weak incentive to follow the specified message strategy m_t. But then the principal and agent t have no profitable deviation from e_t, μ_t, or s_t, since continuation play is exactly as in σ^*. So this strategy profile is an equilibrium of the no-extortion game, as desired. ■

Since agents are indifferent among messages, they are always willing to follow through on their threats. If they do, then the resulting mapping from transfer to message is identical to the corresponding mapping in the extortion game, leading to identical equilibrium outcomes. The only complication to this argument arises in the extortion game with bilateral relationships, since an agent's payoff in the coordination game can potentially respond to his message. However, we can always find an equilibrium in which agents are punished in the bilateral relationship if they deviate from their threats, in which case agents are willing to follow through on those protocols.

Since agents are indifferent among their messages in the extortion game, even a small intrinsic preference for following through on threats is enough to replicate Proposition 2. To make this point, we consider the game with ϵ-compliance preferences, which is identical to the no-extortion game except that each agent t earns an additional $\epsilon > 0$ payoff for choosing $m_t = \mu_t(s_t)$. This small preference for complying with the threat is enough to lead to the complete collapse of effort in equilibrium.

Proposition 8 *For any $\epsilon > 0$, every equilibrium in the game with ϵ-compliance preferences has $e_t = s_t = 0$ in every $t \geq 0$.*

Proof: Fix $\epsilon > 0$. Consider an equilibrium of the game with ϵ-compliance preferences. Since agent t is otherwise indifferent among messages, he sends $m_t = \mu_t(s_t)$ in every equilibrium. For each μ_t, the equilibrium mapping from s_t to m_t is identical to that of an equilibrium of the extortion game, from which the result follows. ■

Even a small preference for following the threat is enough to break agents' indifference across messages and so replicate our impossibility result. We could apply a similar argument in the extortion game with either effort signals or transfer signals to prove that equilibrium outcomes are similarly equivalent. In contrast, such an equivalence does not hold in the extortion game with bilateral relationships, since the bilateral relationship can be used to deter agents from following their threats if $\epsilon > 0$ is small.

Proposition 8 assumes that agents prefer to "keep their word" by acting according to their threats. Other types of intrinsic preferences could lead to different equilibrium outcomes, including equilibria with strictly positive effort. To illustrate this point, suppose that each agent t instead prefers to "tell the truth," in the sense that he receives an extra $\epsilon > 0$ utility if (i) he sends $m_t = C$ and no deviation occurred in period t, or (ii) he sends $m_t = D$ and a deviation did occur. It is straightforward to show that intrinsic preferences of this sort are enough to restore cooperation to the game level from Proposition 1. Note, however, that agents who prefer to tell the truth earn lower utility than those who can extort the principal, since the former must exert effort to earn a transfer while the latter can shirk. Consequently, if an agent could develop a reputation with the principal (unobserved by other agents), then he would prefer to have a reputation for extortion rather than for telling the truth. By the same logic, organizations that rely on coordinated punishments risk attracting exactly those agents who are most willing to make extortionary threats.

Propositions 7 and 8 suggest two reasons why commitment is a relatively mild assumption in our setting. Fundamentally, however, we introduce commitment for a more applied reason: the threat of extortion features prominently in each of our applications, and studying extortion requires a setting in which agents can make action-contingent threats even after they deviate. The threat, or something like it, is therefore necessary to study extortion and identify new ways to encourage cooperation.

C Online Appendix: Communication by the Principal

Communication among the agents lies at the heart of our analysis. This appendix explores alternative assumptions about communication. In Online Appendix C.1, we show that extortion remains a problem even if the principal can send a public message at the end of each period. Intuitively, if the principal could lessen her punishment by reporting extortion, then she would always do so regardless of whether or not extortion actually occurred. Online Appendix C.2 then shows that extortion *can* be eliminated if the principal can commit to threats as a function of each period's transfer, provided that she makes her threat *weakly before* the agent makes his threat. This positive result should be interpreted with skepticism, however, since unlike the agents, the principal sometimes has an incentive to deviate from her threat.[9] Finally, once the principal pays an extorting agent, that payment is sunk and so the agent has an incentive to extort again. Online Appendix C.3 explores cooperation when agents have multiple opportunities to extort, with the conclusion that extortion has similar effects on equilibrium outcomes in that setting.

C.1 The Principal Can Send Messages

Let M_p be the set of messages for the principal, and m_p a typical message. In each period $t \geq 0$, the principal chooses a message $m_{p,t}$ in each period $t \geq 0$, and this message is publicly observed. We consider two different stage games: the principal might either choose $m_{p,t} \in M_p$ before or after agent t chooses m_t. If the principal chooses $m_{p,t}$ before m_t is realized, we assume that μ_t is a function of s_t only (and so doesn't depend on $m_{p,t}$).

The principal talks after agent t. Consider some period t. We let $\pi(m, m_p)$ be the principal's continuation payoff if (m, m_p) realizes. Given agent t's message m, the principal always chooses m_p to maximize $\pi(m, m_p)$. We let $\pi(m) := \max_{m_p} \pi(m, m_p)$, so $\pi(m)$ is the

[9] That is, unlike its role for agents, commitment forces the principal to send messages that are *ex post* suboptimal. Hence, allowing the principal to commit does *not* refine the equilibrium set of the game without commitment.

principal's continuation payoff after agent t's message m. We let $\overline{\Pi}$ and $\underline{\Pi}$ be the highest and lowest continuation payoffs that agent t's message can induce. Then, incentive constraints are identical to the extortion game (i.e., Proposition 2). The principal's message does not mitigate extortion at all, so our impossibility result still holds.

Proposition 9 *Suppose that in each period t the principal sends $m_p \in M_p$ after agent t sends m. The principal-optimal equilibrium is outcome-equivalent to that in Proposition 2.*

The principal talks before agent t. Consider some period t. Define $\pi(m_p, m)$ as the principal's continuation payoff if $m_t = m$ and $m_{p,t} = m_p$. Once the principal chooses s_t, she knows $m_t = \mu_t(s_t)$. The principal therefore chooses $m_{p,t}$ to maximize her continuation payoff given agent t's message.[10] The same argument as in the previous case applies, so every equilibrium involves zero effort in each period.

C.2 The Principal Can Make Threats

In this appendix, we modify the extortion game by allowing the principal to choose a threat at the same time as each agent. We first show that Proposition 1 holds in this game, which means that allowing the principal to commit to messages as a function of transfers eliminates extortion. We then give two reasons why this result should be treated with skepticism.

Formally, suppose that in each $t \geq 0$, the principal chooses a threat $\nu_t : \mathbb{R} \to M$ at the same time that agent t chooses e_t and μ_t. At the end of t, message $m_t^P = \nu_t(s_t)$ is realized and publicly observed (along with agent t's message m_t). We can adapt the proof of Proposition 1 to show that the principal can earn no more than $e^* - c(e^*)$ in this game, where e^* is defined as in Proposition 1. It suffices to construct an equilibrium in which she earns that payoff.

[10] This intuition would not change if agents could commit to a mixture over M, in which case the principal would choose $m_{p,t}$ to maximize her continuation payoff given the mixture. The key is that agent t can use her message to implement the same punishment regardless of whether he works or shirks.

Consider the following strategy profile. Play starts in the cooperation phase. In this phase,

$$\nu_t(s_t) = \mu_t(s_t) = \begin{cases} C & s_t \geq c(e^*) \\ D & \text{otherwise} \end{cases}$$

and $e_t = e^*$. If neither player deviates, then $s_t = c(e^*)$; if only agent t deviates, then $s_t = 0$; if the principal or both players deviate, then the principal best-responds given the threats. The game stays in the cooperative phase if $m_t = m_t^P = C$. Otherwise, it switches to the punishment phase with probability $\gamma \in [0, 1]$. In the punishment phase, agents exert no effort and the principal pays no transfers.

Choosing γ to solve

$$c(e^*) = \frac{\delta}{1-\delta}\gamma\left(e^* - c(e^*)\right) \tag{20}$$

implies that the principal is willing to pay $s_t = c(e^*)$ on the equilibrium path. If agent t deviates, then the principal's continuation payoff cannot exceed $e^* - c(e^*)$ if she pays $s_t = c(e^*)$ and equals $(1-\gamma)(e^* - c(e^*))$ if she pays any other amount. Condition (20) implies that she is willing to pay $s_t = 0$ in that case. Agent t therefore has no profitable deviation from e_t or μ_t. The principal has no profitable deviation from ν_t, since given μ_t, she earns no more than $e^* - c(e^*)$ for paying $s_t = c(e^*)$ and no more than $(1-\gamma)(e^* - c(e^*))$ for paying any other amount. This strategy profile is therefore an equilibrium. It is principal-optimal because it maximizes total equilibrium surplus and gives all of that surplus to the principal.

This argument shows that allowing the principal to commit to a threat eliminates extortion. Essentially, the principal's and each agent's threats can be used to "cross-check" one another. If the principal is punished whenever messages disagree, then agents cannot extort any *smaller* amount than the amount that the principal pays a hard-working agent on-path. As in the proof of Proposition 4, the principal can then be made indifferent between paying $s_t = c(e^*)$ and $s_t = 0$, so that she is willing to pay a hard-working agent but not one that shirks.

While allowing the principal to commit to a threat can in principle restore cooperation, this result should be treated with skepticism for two reasons. First, while agents are indifferent across messages, the principal is not. Indeed, appendix C.1 shows that she has a strict incentive to send the message that maximizes her continuation payoff. Commitment therefore forces the principal to send messages that she strictly prefers not to send, which stands in contrast to the agents, for whom commitment simply breaks indifference across messages. Consequently, we cannot treat the principal's threat as an equilibrium refinement; no analogue to Proposition 7 exists for the game with principal commitment.

Second, as appendix C.1 illustrates, this result requires the principal to choose ν_t *(weakly) before* agent t chooses μ_t and e_t. If agent t chooses μ_t first, then he can shirk and extort the principal, in which case her unique best-response is to pay that agent and then send a message that guarantees a high continuation payoff. If the principal chooses ν_t before agent t chooses μ_t, in contrast, then we can slightly modify the equilibrium construction above to show that a version of Proposition 1 holds. The conclusion that principal commitment eliminates extortion therefore depends on a particular assumption about *when* each player makes threats.

C.3 Each Agent has Multiple Extortion Opportunities

This section considers equilibria if agents have multiple opportunities to make threats in the extortion game. Once the principal gives in to an extortion attempt, an agent has every incentive to repeat the same threat in the hopes of extracting yet more money. In equilibrium, the principal should anticipate that each payment might not be the final one. What is the effect on equilibrium cooperation?

We introduce the **extortion game with repeated threats** to address this question. In each period $t \in \{0, 1, ...\}$, the principal and agent t play the following stage game:

1. Agent t chooses $e_t \in \mathbb{R}_+$.

2. The following payment subgame is played repeatedly. At the end of each repetition, the stage game moves to the next stage with probability $\rho \in (0,1)$, and otherwise the payment subgame repeats. In repetition $k \in \{1, 2, ...\}$ of the payment subgame:

   (a) Agent t chooses $\mu_t^k : \mathbb{R} \rightarrow \mathcal{M}$;

   (b) The principal chooses a transfer $s_t^k \in \mathbb{R}_+$.

3. Let $K \in \mathbb{R}$ be the final iteration of the payment subgame. Then $s_t = \sum_{k=1}^K s_t^k$ and $m_t = \mu_t^K(s_t^K)$.

The principal and agent t's payoffs are identical to the extortion game. In particular, the principal does not discount between iterations of the payment subgame.

This model assumes that agents can threaten the principal an unknown number of times, and that only the message associated with the final threat is observed by other agents. If each agent could threaten the principal a known, finite number of times, then only their final threats would matter in equilibrium, so the analysis from the baseline extortion model would apply.

We show that our results from the extortion game hold even if each agent can make an uncertain number of threats, provided that the probability of being able to make one additional threat, $1 - \rho$, is not too large. To prove this result, we characterize the amount that an agent can extort as a function of his leverage.

Proposition 10 *Consider an equilibrium of the payment subgame, and let $\bar{\Pi}$ and $\underline{\Pi}$ equal the best-case and worst-case principal payoffs, respectively, with corresponding messages $\bar{m}$ and $\underline{m}$. If $\rho > \frac{1}{2}$, then*

$$\mathbb{E}\left[s_t\right] = \frac{\delta}{1-\delta}\left(\bar{\Pi} - \underline{\Pi}\right)$$

in any equilibrium.

Proof of Proposition 10

Let U_M and U_m be the agent's largest an smallest equilibrium payoffs in the payment subgame. Our goal is to show that for any $\rho > \frac{1}{2}$, $U_M = U_m = \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}) \equiv L$.

The following equilibrium strategy gives agent t a payoff of L: in each k,

$$\mu_t^k = \begin{cases} \bar{m} & s_t^k = \rho L \\ \underline{m} & \text{otherwise.} \end{cases}$$

The principal pays $s_t^k = \rho L$. The probability of the payment subgame surviving to iteration k equals $(1-\rho)^k$, so this strategy profile gives the agent an expected payoff

$$U_M = \sum_{k=0}^{\infty} (1-\rho)^k \rho L = L.$$

The principal's expected payoff equals $\frac{\delta}{1-\delta}\underline{\Pi}$. Continuation play is independent of s_t^k, so the principal is willing to pay ρL, because

$$-\rho L + \rho \bar{\Pi} + \rho \underline{\Pi} = \underline{\Pi}.$$

Agent t has no profitable deviation from μ_t^k, since the principal would be unwilling to pay any amount larger than ρL. Thus, this strategy is an equilibrium. Moreover, $U_M \leq L$, because the principal cannot earn a payoff lower than $\frac{\delta}{1-\delta}\underline{\Pi}$ in equilibrium, and total surplus cannot exceed $\frac{\delta}{1-\delta}\bar{\Pi}$.

Now, we bound U_m from below. The principal's minimum equilibrium payoff equals $\frac{\delta}{1-\delta}\underline{\Pi}$; let $\frac{\delta}{1-\delta}\Pi_M$ equal her maximum equilibrium payoff. Then, the principal's **unique** best response to

$$\mu_t^k = \begin{cases} \bar{m} & s_t^k = s \\ \underline{m} & \text{otherwise} \end{cases} \tag{21}$$

equals $s_t^k = s$, so long as

$$-s + \rho\frac{\delta}{1-\delta}\bar{\Pi} + (1-\rho)\frac{\delta}{1-\delta}\underline{\Pi} > \rho\frac{\delta}{1-\delta}\underline{\Pi} + (1-\rho)\frac{\delta}{1-\delta}\Pi_M,$$

or

$$s < (1-2\rho)\frac{\delta}{1-\delta}\underline{\Pi} + \rho\frac{\delta}{1-\delta}\bar{\Pi} - (1-\rho)\frac{\delta}{1-\delta}\Pi_M.$$

Let s_M equal the *supremum* transfer that satisfies this constraint, with $s_M = 0$ if no transfer does. Then,

$$\Pi_M \leq \frac{\delta}{1-\delta}\bar{\Pi} - \sum_{k=0}^{\infty}(1-\rho)^k s_M,$$

since if agent t earns less than $\sum_{k=0}^{\infty}(1-\rho)^k s_M = \frac{s_M}{\rho}$, he can profitably deviate to (21) in each k, with $s = s_M - \epsilon$ for $\epsilon > 0$ arbitrarily small.

By definition of s_M, we must have

$$s_M = \max\left\{0, (1-2\rho)\frac{\delta}{1-\delta}\underline{\Pi} + \rho\frac{\delta}{1-\delta}\bar{\Pi} - (1-\rho)\left(\frac{\delta}{1-\delta}\bar{\Pi} - \frac{s_M}{\rho}\right)\right\}.$$

Simplifying, we have

$$s_M = \max\left\{0, (2\rho-1)\frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}) + \frac{1-\rho}{\rho}s_M\right\}.$$

For $\rho > \frac{1}{2}$, the right-hand side of this equality is strictly positive. In that case, we can gather terms to yield

$$\frac{2\rho-1}{\rho}s_M = (2\rho-1)\frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}).$$

Cancelling $2\rho - 1$ from both sides of this expression yields

$$s_M = \frac{\delta}{1-\delta}\rho L,$$

in which case $U_m \geq \sum_{k=0}^{\infty}(1-\rho)^k \rho L = L$.

We conclude that if $\rho > \frac{1}{2}$, agent t's unique equilibrium payoff equals L, so $\mathbb{E}[s_t] = L$, as desired.$\blacksquare$

What happens if $\rho < \frac{1}{2}$?

The payment subgame resembles a repeated game, where the probability of continuing to another iteration, $1 - \rho$, corresponds to the discount factor. This subgame is also positive-sum; feasible total surplus can be as low as $\frac{\delta}{1-\delta}\underline{\Pi}$ or as high as $\frac{\delta}{1-\delta}\bar{\Pi}$. Consequently, for $\rho < \frac{1}{2}$, we can use repeated-game incentives to deter agent t from extorting. One way to construct these incentives is familiar from Section 5: the principal is punished after she gives in to an extortion attempt. The principal therefore refuses to pay anything following a deviation, so the agent refrains from extortion.

This equilibrium construction might not be possible in practice, since it requires extortion attempts to be structured in a way that facilitates the use of repeated-game style incentives. Nevertheless, we can construct this kind of equilibrium when $\rho < \frac{1}{2}$, so it represents an alternative potential remedy to extortion.

D Online Appendix: Long-run Agents

D.1 A Result with long-run Agents

D.1.1 Model, Result, and Discussion

Consider a repeated game with a single principal and N agents with a shared discount factor $\delta \in [0, 1)$. In each period, the following stage game is played:

1. Exactly one agent is publicly selected to be active. For each agent $i \in \{1, ..., N\}$, let $x_{i,t} \in \{0, 1\}$ be the indicator function for agent i being selected. Let $\Pr\{x_{i,t} = 1\} = \rho_i$, where $\sum_i \rho_i = 1$.

2. The active agent chooses $e_t \in \mathbb{R}_+$ and $\mu_t : \mathbb{R} \to M$, which are observed only by the principal and the active agent.

3. The principal and the active agent exchange transfers, with resulting net transfer to the active agent $s_t \in \mathbb{R}$. These transfers are observed only by the principal and the active agent.

4. The message $m_t = \mu_t(s_t)$ is realized and publicly observed.

The principal's and agent i's payoffs in each period t are $\pi_t = e_t - s_t$ and $u_{i,t} = x_{i,t}(s_t - c(e_t))$, respectively, with corresponding expected discounted payoffs $\Pi_t = \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)(e_t - s_t)$ and $U_{i,t} = \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)x_{i,t}(s_t - c(e_{i,t}))$. Our solution concept is plain Perfect Bayesian Equilibrium with one additional restriction: at any history h^t such that agent i has observed a deviation, we require that $\mathbb{E}[U_{i,t}|h^t] \geqslant 0$. This restriction rules out pathological off-path behavior that might arise from the fact that an agent's beliefs about the history are essentially arbitrary once he observes a deviation.[11] We also restrict attention to equilibria in pure strategies to simplify agents' beliefs on the equilibrium path.

Proposition 11 *Let e_i^* be the maximum effort attainable in any pure-strategy Perfect Bayesian Equilibrium. Letting $s_i^* \equiv \min\{e_i^*, e^{FB}\} - c(\min\{e_i^*, e^{FB}\})$, $e_1^*, e_2^*, ..., e_N^*$ must satisfy the system of inequalities*

$$(1-\delta)c(e_i^*) \leqslant 2\delta\rho_i s_i^* + \frac{2\rho_i\delta}{1-(1-\rho_i)\delta}\sum_{j\neq i}\rho_j s_j^*. \tag{22}$$

It is instructive to compare the right-hand side of (22) to the condition $c(e^*) \leq 3(H-L)$ from Proposition 5. To translate between settings, note that the total surplus created by the principal's future interactions with agent i equals $\delta\rho_i s_i^*$, which corresponds to $2(H-L)$ in Proposition 5. In Proposition 5, the principal earns an additional $H-L$ if she refuses to pay an agent who has shirked. In the game with long-run agents, the principal can be given

[11] This condition is trivially satisfied in any equilibrium that is recursive. It is needed here because this game has private monitoring, which means that equilibria are not necessarily recursive.

the *entire* continuation surplus from her relationship with agent i, which accounts for the second $\delta \rho_i s_i^*$ in the right-hand side of (22).

The second term on the right-hand side of (22) represents a new force for cooperation that is not present in Proposition 5. Since each agent i chooses a new threat whenever he is active, he essentially commits to his messages *only until he next interacts with the principal again.* An agent might therefore use his future messages to reveal that he has extorted the principal in equilibrium. However, he cannot do so until the next time that he is active, so this term shrinks to zero as $\rho_i \to 0$.

An immediate corollary of Proposition 11 is that, as the probability that an agent interacts with the principal ρ_i approaches zero, that agent's maximum equilibrium effort does too. This implication is similar to our main takeaway from Proposition 5: the strength of each agent's bilateral relationship limits the severity of the coordinated punishments available to him. This result relies on the fact that agents can send messages only when they are active. We can interpret this assumption as the natural extension of our commitment assumption to a setting with long-run agents; indeed, a result identical to Proposition 11 would hold if agents could communicate in every period but whenever an agent is active, he commits to a *sequence* of messages in each period until he is again active.

D.1.2 Proof of Proposition 11

For each agent $j \in \{1, ..., 2\}$, let e_j^* be the maximum effort that can be attained in any period of any equilibrium. Consider a history h^t right after agent i is chosen to be the active agent in period t. Define four different expectations of Π_{t+1} that follow four different outcomes:

1. $\overline{\Pi}^*$ if no player deviates, with corresponding message $\overline{m}$;

2. $\underline{\Pi}^*$ if the principal deviates but the active agent does not, with corresponding message $\underline{m}$;

3. $\overline{\Pi}^{HU}$ if the active agent deviates and $m_t = \overline{m}$;

4. $\underline{\Pi}^{HU}$ if the active agent deviates and $m_t = \underline{m}$.

We identify necessary conditions for effort e to be attained in equilibrium.

First, the principal must be willing to pay s^* if the active agent does not deviate, which requires

$$s^* \leqslant \frac{\delta}{1-\delta}\left(\overline{\Pi}^* - \underline{\Pi}^*\right). \tag{23}$$

Second, the active agent i must be willing to choose effort e and the equilibrium threat μ. Agent i can always deviate by choosing $e_t = 0$ and

$$\mu_t = \begin{cases} \underline{m} & s_t < \hat{s} \\ \overline{m} & \text{otherwise} \end{cases}$$

for some $\hat{s} \geqslant 0$. Following this deviation, the principal's unique best response is to pay $\hat{s}$ so long as

$$-\hat{s} + \frac{\delta}{1-\delta}\overline{\Pi}^{HU} > \frac{\delta}{1-\delta}\underline{\Pi}^{HU},$$

since the principal can earn no less than $\overline{\Pi}^{HU}$ in the continuation game if $m_t = \overline{m}$ and no more than $\underline{\Pi}^{HU}$ if $m_t = \underline{m}$. Therefore, agent i has no profitable deviation of this form only if

$$s^* - c(e) + \frac{\delta}{1-\delta}\overline{U}_i^* \geqslant \max\left\{0, \frac{\delta}{1-\delta}\left(\overline{\Pi}^{HU} - \underline{\Pi}^{HU}\right)\right\}, \tag{24}$$

where $\overline{U}_i^*$ is the agent's expectations about her continuation payoff at the history that yields principal payoff $\overline{\Pi}_i^*$.

Combining (23) and (24) yields the following necessary condition for effort e to be part of equilibrium:

$$c(e) \leqslant \frac{\delta}{1-\delta}\left(\overline{U}_i^* + \overline{\Pi}^* - \underline{\Pi}^*\right) - \max\left\{0, \frac{\delta}{1-\delta}\left(\overline{\Pi}^{HU} - \underline{\Pi}^{HU}\right)\right\} \tag{25}$$

Our next goal is to connect (23) and (24) by studying the relationship between $\overline{U}_i^* + \overline{\Pi}^* -$

$\underline{\Pi}^*$ and $\overline{\Pi}^{HU} - \underline{\Pi}^{HU}$. We do so by bounding $\overline{U}_i^* + \overline{\Pi}^* - \overline{\Pi}^{HU}$ from above and $\underline{\Pi}^* - \underline{\Pi}^{HU}$ from below.

Fix two histories h^{t+1} and $\hat{h}^{t+1}$ at the start of period $t+1$ such that agent i can distinguish h^{t+1} from $\hat{h}^{t+1}$ but no other agents can. For $t' \geqslant t+1$, we will use the notation $h^{t'}$ and $\hat{h}^{t'}$to represent successor histories to h^{t+1} and $\hat{h}^{t+1}$, respectively. At history $\hat{h}^{t+1}$, the principal can always play the following strategy:

1. At any history $\hat{h}^{t'}$ that the active agent believes is consistent with h^{t+1}, play as in the corresponding successor history to h^{t+1};

2. At any other history, choose $s_t = 0$.

Under this strategy, each agent $j \neq i$ learns that the history is inconsistent with h^{t+1} only when agent i sends a message that is inconsistent with play following h^{t+1}. In a pure-strategy equilibrium, all agents learn this fact at the same time. For each $t' \geqslant t+1$, denote

$$\hat{\mathcal{B}}^{t'} = \left\{ \hat{h}^{t'} | \text{Agents } j \neq i \text{ learn that the history is inconsistent with } h^{t+1} \text{ in period } t'-1, \right. \\ \left. \text{but not before} \right\}.$$

Where $\hat{\mathcal{B}}^{\infty}$ denotes the event that agents $j \neq i$ never learn that the history is inconsistent with h^{t+1}. Note that these events collectively partition the set of histories following $\hat{h}^{t+1}$. We can define an analogous collection of sets for the event that agents $j \neq i$ learn that the history is inconsistent with $\hat{h}^{t+1}$. We denote this analogous collection $\mathcal{B}^{t'}$.

For each agent $j \in \{1, ..., N\}$, define $\pi_{j,t} = x_{j,t}(e_t - s_t)$ and $\pi_{-j,t} = \sum_{k \neq j} x_{k,t}(e_t - s_t)$ as the principal's payoff from agent j and from all other agents, respectively. Define $\Pi_{j,t} = \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)\pi_{j,t'}$ and $\Pi_{-j,t} = \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)\pi_{-j,t'}$. Because $\left\{ \hat{\mathcal{B}}^{\tilde{t}} \right\}_{\tilde{t}=t+1}^{t'}$ partitions the histories of length t' following $\hat{h}^{t+1}$,

$$\mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] = \sum_{t'=t+1}^{\infty}(1-\delta)\delta^{t'-t-1}\left(\mathbb{E}\left[\pi_{i,t'}|\hat{h}^{t+1}\right] + \sum_{\tilde{t}=t+1}^{t'}\mathbb{E}\left[\pi_{-i,t'}|\hat{h}^{t+1},\hat{\mathcal{B}}^{\tilde{t}}\right]\Pr\left\{\hat{\mathcal{B}}^{\tilde{t}}\right\}\right). \tag{26}$$

The right-hand side of (26) is absolutely convergent, so we can rearrange the order of summation to yield

$$\mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] = \begin{array}{c} \sum_{t'=t+1}^{\infty}(1-\delta)\delta^{t'-t-1}\mathbb{E}\left[\pi_{i,t'}|\hat{h}^{t+1}\right] + \\ \sum_{\tilde{t}=t+1}^{\infty}\left(\sum_{t'=t+1}^{\tilde{t}-1}(1-\delta)\delta^{t'-t-1}\mathbb{E}\left[\pi_{-i,t'}|\hat{\mathcal{B}}^{\tilde{t}}\right] + \delta^{\tilde{t}-t-1}\mathbb{E}\left[\Pi_{-i,\tilde{t}}|\hat{\mathcal{B}}^{\tilde{t}}\right]\right)\Pr\left\{\hat{\mathcal{B}}^{\tilde{t}}\right\}. \end{array} \tag{27}$$

Under the principal's strategy specified above, the principal and agents $j \neq i$ act identically until those agents learn of a deviation. Therefore, for any $t' < \tilde{t}$,

$$\mathbb{E}\left[\pi_{-i,t'}|\mathcal{B}^{\tilde{t}}\right]\Pr\left\{\mathcal{B}^{\tilde{t}}\right\} = \mathbb{E}\left[\pi_{-i,t'}|\hat{\mathcal{B}}^{\tilde{t}}\right]\Pr\left\{\hat{\mathcal{B}}^{\tilde{t}}\right\}.$$

Moreover, for any $\tilde{t}$, $\Pr\left\{\mathcal{B}^{\tilde{t}}\right\} = \Pr\left\{\hat{\mathcal{B}}^{\tilde{t}}\right\}$, since any message that distinguish h^{t+1} from $\hat{h}^{t+1}$ must also distinguish $\hat{h}^{t+1}$ from h^{t+1}.

Now, $\mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right]$ is bounded below by the principal's payoff from the strategy specified above. Therefore, we can use (27) to bound the difference

$$\begin{array}{c} \mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] - \mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] \leqslant \\ \sum_{t'=t+1}^{\infty}\delta^{t'-t-1}(1-\delta)\left(\mathbb{E}\left[\pi_{i,t'}|h^{t+1}\right] - \mathbb{E}\left[\pi_{i,t'}|\hat{h}^{t+1}\right]\right) + \\ \sum_{\tilde{t}=t+1}^{\infty}\delta^{\tilde{t}-t-1}\left(\mathbb{E}\left[\Pi_{-i,\tilde{t}}|\mathcal{B}^{\tilde{t}}\right] - \mathbb{E}\left[\Pi_{-i,\tilde{t}}|\hat{\mathcal{B}}^{\tilde{t}}\right]\right)\Pr\left\{\mathcal{B}^{\tilde{t}}\right\} \end{array} \tag{28}$$

Under the specified strategy, $\mathbb{E}\left[\Pi_{-i,\tilde{t}}|\hat{\mathcal{B}}^{\tilde{t}}\right] \geqslant 0$ because the principal pays no transfer to an agent j who knows that the history is inconsistent with h^{t+1}, with $\mathbb{E}\left[\pi_{i,t'}|\hat{h}^{t+1}\right] \geqslant 0$ for a similar reason. A necessary condition for (28) is therefore

$$\begin{gathered}\mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] - \mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] \leqslant \\ \textstyle\sum_{t'=t+1}^{\infty} \delta^{t'-t-1}\left((1-\delta)\mathbb{E}\left[\pi_{i,t'}|h^{t+1}\right] + \mathbb{E}\left[\Pi_{-i,t'}|\mathcal{B}^{t'}\right]\Pr\left\{\mathcal{B}^{t'}\right\}\right)\end{gathered} \tag{29}$$

Suppose that h^{t+1} is the on-path history such that $\mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] = \overline{\Pi}^{*}$. In a pure-strategy equilibrium, agents correctly infer the true history on the equilibrium path, which means that they must earn nonnegative utility. Consequently, the principal earns no more than total continuation surplus, so (29) requires

$$\mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] - \mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] \leqslant \sum_{t'=t+1} \delta^{t'-t-1}\left((1-\delta)\rho_i s_i^* + \sum_{j\neq i}\rho_j s_j^* \Pr\left\{\mathcal{B}^{t'}\right\}\right). \tag{30}$$

Note than an identical bound holds for the expression $\overline{U}_i^* + \overline{\Pi}^* - \underline{\Pi}^*$ because agents $j \neq i$ earn nonnegative continuation utilities on the equilibrium path. If h^{t+1} is instead the history such that $\mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] = \underline{\Pi}^{HU}$, then agents have observed $\underline{m}$ and so know that play is off-path. Our equilibrium restriction requires their utilities to be nonnegative at such a history, so (30) again holds.

Since $s_j^* \geqslant 0$, the right-hand side of (30) is maximized by having the event $\mathcal{B}^{\tilde{t}}$ happen as early as possible. The earliest it can occur is the next time that agent i is the active agent, since agent i can send a message only when he is active. Agent i is active for the first time since period t in period t' with probability $(1-\rho_i)^{t'-t-1}\rho_i$, so (30) requires

$$\begin{aligned}\mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] - \mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] &\leqslant \textstyle\sum_{t'=t+1} \delta^{t'-t-1}\left((1-\delta)\rho_i s_i^* + (1-\rho_i)^{t'-t-1}\rho_i \sum_{j=1}^{N}\rho_j s_j^*\right) \\ &= \rho_i s_i^* + \tfrac{\rho_i}{1-(1-\rho_i)\delta}\textstyle\sum_{j\neq i}\rho_j s_j^*.\end{aligned} \tag{31}$$

As argued above, an identical bound holds for $\mathbb{E}\left[U_{i,t+1} + \Pi_{t+1}|h^{t+1}\right] - \mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right]$.

From (31), we conclude that

$$\overline{U}_i^* + \overline{\Pi}^* - \underline{\Pi}^* \leqslant \overline{\Pi}^{HU} - \underline{\Pi}^{HU} + 2\rho_i s_i^* + \frac{2\rho_i}{1-(1-\rho_i)\delta}\sum_{j\neq i}\rho_j s_j^*.$$

A necessary condition for (25) to hold is therefore

$$c(e) \leqslant \left(\begin{array}{c} \frac{\delta}{1-\delta}\left(\overline{\Pi}^{HU} - \underline{\Pi}^{HU} + 2\rho_i s_i^* + \frac{2\rho_i}{1-(1-\rho_i)\delta}\sum_{j\neq i}\rho_j s_j^*\right) - \\ \max\left\{0, \frac{\delta}{1-\delta}\left(\overline{\Pi}^{HU} - \underline{\Pi}^{HU}\right)\right\} \end{array} \right).$$

The right-hand side of this condition is maximized by $\overline{\Pi}^{HU} - \underline{\Pi}^{HU} = 0$, in which case

$$(1-\delta)c(e) \leqslant 2\rho_i s_i^* + \frac{2\rho_i}{1-(1-\rho_i)\delta}\sum_{j\neq i}\rho_j s_j^*,$$

as desired. ■