The Use and Misuse of Coordinated Punishments

Daniel Barron and Yingni Guo*

March 25, 2019

Abstract

Communication facilitates cooperation by ensuring that deviators are collectively punished. We explore how players might misuse equilibrium messages to threaten one another, and we identify conditions under which communication improves cooperation despite such threats. In our model, a principal plays trust games with a sequence of short-run agents who communicate with each other. A shirking agent can demand pay by threatening to report that the principal deviated. We show how these threats can destroy cooperation. However, some cooperation is restored if players observe public signals of efforts or transfers, or if the principal has a bilateral relationship with each agent.

*Barron: Northwestern University, Kellogg School of Management, Evanston IL 60208; email: d-barron@kellogg.northwestern.edu. Guo: Northwestern Unviersity, Economics Department, Evanston IL 60208; email: yingni.guo@northwestern.edu. The authors would like to thank Charles Angelucci, Nemanja Antic, 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, Arijit Mukherjee, Jacopo Perego, Michael Powell, Luis Rayo, Jonah Rockoff, Mark Satterthwaite, Takuo Sugaya, Jeroen Swinkels, Joel Watson, and audiences at Arizona State University, University of Texas - Austin, University of Waterloo, Michigan State University, University of Michigan, and the 2018 Junior Workshop on Organizational Economics. 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 organizations and markets, however, cooperation is not automatic: participants typically refrain from opportunistic behavior only because they expect such behavior to be punished (Malcomson (2013)). Communication plays an essential role in coordinating these punishments by allowing news of misbehavior to spread far beyond those who directly observe that misbehavior. So, for example, workers use the threat of collective punishments to make sure that managers pay promised rewards (Levin (2002)); guilds and industry associations encourage fair dealing between both members and nonmembers by threatening boycotts (Greif et al. (1994); Bernstein (2015)); communities protect public resources by threatening noncontributors with communal sanctions (Ostrom (1990)); and online marketplaces publicize buyer reviews so that poor service will hurt future sales (Hörner and Lambert (2018)).

Communication is powerful in these settings because it can be used to coordinate punishments of those who have reneged on their promises. But individuals armed with this power also face a grave temptation, since they can demand concessions from their partners by threatening to *falsely* spread word of misbehavior. We will show that these threats have the potential to undermine coordinated punishments and destroy cooperation. Despite their prevalence and potentially dire consequences, however, the constraints that these threats impose on cooperative relationships remain understudied.

In this paper, we explore how rent-seeking individuals can misuse messages that are designed to encourage cooperation, and we identify remedies that facilitate cooperation in the face of such misuse. We develop a principal-agents model in which cooperation depends on communication among the agents. This model allows each agent to shirk and threaten to make false reports unless the principal takes a desired action, a deviation that we will call **extortion.** We first prove a stark impossibility result: the possibility of extortion totally undermines cooperation, since any attempt to coordinate punishments simply becomes an opportunity for agents to extract surplus without exerting effort. We then build on this

impossibility result to identify ways that coordinated punishments can improve cooperation despite these threats. Collectively, our results suggest that coordinated punishments are susceptible to misuse, but they also identify organizational designs that limit that misuse and so restore cooperation.

To illustrate this kind of misuse, consider the relationship between a manager (of, say, a manufacturing plant) and her workers. These relationships are rarely governed by formal contracts alone; instead, workers are willing to strive for excellent performance only if they trust the manager to reward their efforts (Gibbons and Henderson (2013)). In this context, an institution that allows workers to collectively punish a misbehaving manager – whether that institution is a traditional union (Freeman and Medoff (1979)) or an online job review platform like Glassdoor.com – can encourage cooperation between the manager and workers and so enable highly productive outcomes. However, if such institutions are not carefully designed, then workers may face the temptation to subvert them in pursuit of private gain. In some cases, this subversion is shockingly overt, as in the recent conviction of an Illinois union official who extorted personal bribes from businesses by threatening to initiate labor unrest (Meisner and Ruthhart (2017)). Subtler forms of subversion can also have serious consequences. For example, Glass and Langfitt (2015) records how manipulative grievances, filed not in accord with the union's wishes but instead by individual workers hoping to gain leverage over management, led to very low productivity at General Motor's Fremont plant in the 1980s.[1]

The use of coordinated punishments, and their potential for misuse, extend far beyond the factory floor. Firms use industry associations like the Financial Industry Regulatory Authority (FINRA) to share information about employees who provide deficient effort or

[1]Glass and Langfitt (2015) reports that some workers used this leverage to protect themselves from being punished for "shirking" behaviors, including absenteeism and drug use, which were decidedly *not* condoned by the union as a whole. The Fremont plant eventually closed as a result of low productivity, although Toyota later stepped in to reopen it as part of a joint venture with GM. After reopening, Toyota retained almost all of the same unionized workforce but achieved much higher productivity. Consistent with this example, sections 4 and 5 suggest ways to design institutions like unions so that they improve productivity to the benefit of both managers and workers.

engage in unethical behavior. In 2016, however, former employees of Wells Fargo alleged that after they raised concerns about the unfolding scandal at their company, it falsified reports to FINRA in order to have them essentially blacklisted within the financial services industry (Arnold and Smith (2016)). Similar threats are a real concern in online marketplaces and review aggregators. For instance, Klein et al. (2016) shows that extortion was prevalent in the early days of eBay: sellers would extort positive reviews by threatening to leave negative reviews of any buyer that complained, leading to less seller effort and lower buyer satisfaction. To combat similar problems, TripAdvisor explicitly forbids buyers from using the threat of negative reviews to extort firms.[2]

In these examples, messages can both be used to facilitate cooperation and misused to make extortionary threats. We explore the resulting tension using 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 pay a transfer to that agent. Agents observe only their own interaction with the principal but can communicate with other agents at no cost. 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 choose a **communication protocol** at the start of his interaction with the principal. This protocol, which is observed by the principal but not by other agents, associates a message to each possible transfer to that agent. Agents are then committed to communicate according to their protocols.[3]

Agents are willing to exert effort only if they are compensated for doing so, while the principal is willing to compensate an agent only if she would otherwise be punished. Communication is therefore essential for inducing effort, since short-run agents cannot directly punish the principal for reneging. Once armed with messages that can trigger punishments by other agents, however, an agent can shirk and then demand pay using the same threats as an agent who exerted effort. This deviation – an agent shirking and then threatening to

[2]The specific TripAdvisor policy is at https://www.tripadvisor.com/TripAdvisorInsights/w592.

[3]See Wolitzky (2012) and Chassang and Padro i Miquel (2018), which make similar commitment assumptions to study other kinds of action-contingent promises or threats.

say that the principal deviated unless she pays him – is our notion of extortion. Since these threats are enough to induce the principal to compensate an agent for his effort, they are also enough to induce the principal to pay a shirking agent. Consequently, an agent can force the principal to pay him regardless of his effort, which eliminates his incentive to work hard. Section 3 shows that the resulting unique equilibrium outcome entails zero effort.

This impossibility result follows from the observation that a shirking agent can make the same threats, and hence demand essentially the same transfer, as one who works hard. Consequently, anything that drives a wedge between the threats available to a hard-working agent and those available to a shirking agent could lead to positive equilibrium effort. The second half of the paper illustrates this idea with two remedies that create exactly this kind of wedge: public signals and bilateral relationships. These remedies represent concrete ways for organizations to implement coordinated punishments without inviting extortion.

In section 4, we consider **public signals** of either the agents' efforts or the transfers they receive. We interpret these signals as investigations, such as when a union investigates the grievances filed by its members. Public signals allow future agents to condition punishments on more than just previous agents' messages and so tie the threats that an agent can make to his actual interaction with the principal. We show how these types of signals can mitigate extortion and lead to positive equilibrium effort. However, this remedy is not a panacea: if signals are noisy, then the possibility of extortion still reduces equilibrium cooperation.

Section 4.1 considers signals of effort, which can ensure that agents are able to trigger punishments only if they actually exert effort. Perfect effort signals are enough to completely eliminate extortion. Noisy signals, on the other hand, still allow agents to earn some rent by extorting the principal. Hence, agents prefer noisy signals of effort – "flawed investigatory processes" – over perfect signals in order to maximize their rents. Noisy signals also lead to lower equilibrium effort, since the principal is less willing to pay a large transfer today if she anticipates paying rents to future agents. Section 4.2 considers signals of transfers, which in some cases can be used to make the principal indifferent between paying a transfer or not,

so that she is willing to both pay an agent who works hard and not pay an agent who shirks. The downside of this remedy is that it is both fragile (in that it requires the principal to be exactly indifferent) and inefficient (in that the principal is occasionally punished in the resulting equilibrium).

In many applications, agents have multiple opportunities to interact with the principal. Section 5 studies how these future **bilateral relationships** between the principal and each agent can deter extortion and encourage effort. We model these relationships in a reduced-form way by allowing each agent to play a coordination game with the principal at the end of their interaction.[4] As is familiar from the literature on relational contracts (e.g., Levin (2003)), bilateral relationships can be used to directly punish both an agent who shirks and the principal who reneges on a hard-working agent. In addition to these two familiar uses, we show that bilateral relationships can also make extortion less attractive by rewarding the principal for refusing to give in to attempted extortion. That is, bilateral relationships complement coordinated punishments: strong bilateral relationships render extortion less attractive, while weak bilateral relationships lead to potentially lucrative extortionary threats and thus ineffective coordinated punishments.

Without the assumption that agents commit to their communication protocols, we could always select an equilibrium in which agents do not follow through on extortionary threats. Therefore, commitment, or something like it, is a necessary assumption to study the extortionary threats in our applications. In Section 6, we show that the communication protocol can be reinterpreted as an equilibrium refinment in the game without commitment. In other words, we can always find equilibria of the game without commitment such that agents are willing to follow through on their extortionary threats. In this sense, commitment allows agents to specify how they plan to respond to the principal but does not force them to send messages that are *ex post* suboptimal. We also show that with the exception of our analysis of bilateral relationships, our results still hold if we replace commitment with the assumption

[4]Appendix B complements this analysis by studying a setting with truly long-run agents who play a repeated game with the principal.

that agents have a mild intrinsic preference for following their communication protocols.

Related Literature

Our analysis builds on the classic studies that show how institutions can facilitate communication and coordinate punishments (Milgrom et al. (1990), Greif et al. (1994), Dixit (2003)). 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), Ali and Liu (2018)). Within this literature, our paper is closely related to 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. We focus on a different but complementary challenge to coordinated punishments – extortion. More distantly related papers that study communication in repeated games include Awaya and Krishna (2018), which identifies settings in which cooperation requires communication; and Compte (1998) and Kandori and Matsushima (1998), which prove folk theorems in games with private monitoring.

This literature has devoted less attention to how coordinated punishments might be misused. An exception is Bowen et al. (2013), which studies how a community might allow pairs of agents to act on local information. While that paper's setting and analysis both differ from ours, it shares the feature that players might misuse sanctions. In that paper, misuse can be eliminated by delaying communication to the end of each interaction, a remedy that would not work in our setting for two reasons. First, unlike Bowen et al. (2013), our agents communicate about past actions rather than a payoff-relevant state, so our communication already occurs at the end of each interaction. Second, extortion is action-contingent – "pay me *or else* I will lie" – in a way that Bowen et al. (2013)'s misuse is not. The communication protocol is what allows our agents to make these kinds of action-contingent threats.

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 bargaining in

repeated games. Several papers assume that players bargain over either a per-period surplus or continuation play (Baker et al. (2002), Halac (2012, 2015), Goldlucke and Kranz (2017)). The closest papers within this literature are Miller and Watson (2013) and Miller et al. (2018), which define and analyze an equilibrium refinement that essentially allows players to bargain over continuation equilibria. By focusing on communication across agents, our paper analyzes a different friction. In our paper, the principal's "outside option" depends on how messages affect future agents' actions in equilibrium.

More broadly, our framework builds on the relational contracting literature. Much of this literature considers interactions between one principal and one agent (Bull (1987), MacLeod and Malcomson (1989), Baker et al. (1994), Levin (2003)), while we focus on communication across agents. Levin (2002), which studies relational contracts between a principal and multiple agents, is the seminal paper on coordinated punishments within this literature. However, that paper does not consider extortion.

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 must rely 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 relatively ineffective at sustaining cooperation.

Our assumption that agents commit to messages bears a superficial resemblance to the commitment assumption in the persuasion literature (Rayo and Segal (2010), Kamenica and Gentzkow (2011), Lipnowski et al. (2018), Guo and Shmaya (2018)). However, commitment plays a different role in our setting, since it allows agents to make threats rather than reveal information about a payoff-relevant state. As a result, and in contrast to the persuasion lit-

erature, commitment refines the equilibrium set in our model. Our commitment assumption is more closely related to Chassang and Padro i Miquel (2018), which studies how retaliatory threats deter whistleblowing behavior. Unlike that paper, we study how messages *themselves* can be used to make threats. As in our model, Wolitzky (2012) considers a setting in which one side of the interaction can commit to action-contingent messages, but in the context of a firm that can reward its workers by reporting inflated performance measures to the market. In our model, players rely on communication to coordinate sanctions, rather than using it to reveal a payoff-relevant state.

2 Model

We prove our impossibility result for 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 favor-trading game: agent t exerts effort, the principal observes that effort, and the two parties pay one another. This interaction is observed only by the principal and agent t; however, agent t can send a public message at the end of period t. Our key assumption is that before transfers are paid, agent t chooses a *communication protocol,* which is a mapping from the transfer he receives to the message he sends. This communication protocol, which only the principal observes, determines agent t's message as a function of the realized transfer.

Formally, the stage game in period t is:

1. Agent t chooses his effort $e_t \in \mathbb{R}_+$ and a communication protocol $\mu_t : \mathbb{R} \to M$, where M is a large, finite message space.[5] Both e_t and μ_t are observed by the principal but not by any other agent.

2. The principal and agent t simultaneously pay nonnegative transfers to one another,

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

with the net transfer to agent t denoted by $s_t \in \mathbb{R}$.[6] Transfers are observed by these two players but not by any other agent.

3. The message $m_t = \mu_t(s_t)$ is realized and observed by all players.[7]

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, and strictly convex, and satisfies $c(0) = c'(0) = 0$. We assume that there exists a unique first-best effort level, e^{FB}, that solves $c'(e^{FB}) = 1$. 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 in this model, while agents observe only their own interactions with the principal and all public messages. Our solution concept is a Perfect Bayesian Equilibrium (Watson (2016)).[8] Some of our results focus on *principal-optimal equilibria,* which maximize the principal's *ex ante* expected payoff $\mathbb{E}[\Pi_0]$ among all equilibria.

The key ingredient of this model is that agents can commit to their communication protocols, which is what allows a shirking agent to extort the principal. The communication protocol is a transparent way to show that agents have the incentive to extort the principal and explore how extortion affects equilibrium cooperation. Section 6 shows that we can interpret it as an equilibrium refinement of the game without commitment.

Online appendices C and D explore variants of the extortion game in which the principal can communicate or the players can exchange up-front transfers, respectively. We defer further discussion of these variant models until after our impossibility result in Section 3.

If we interpret the principal as a manager and the agents as her workers, then this model lets us analyze how workers might use communication – whether grievances filed with a union

[6] With the exception of section 5, agents have no incentive to pay the principal, so $s_t \geq 0$ in every period t of any equilibrium.

[7] Allowing μ_t to condition on e_t would not change any of our results.

[8] 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.

or negative reviews on a website – to extort concessions from their manager. Of course, the real world is richer than our model: unions typically investigate grievances before acting on them, and individual workers have long-term relationships with the manager. In sections 4 and 5, we show that either of these enrichments can lead to positive effort in equilibrium. Crucially, however, the possibility of extortion still limits equilibrium effort.

3 Threats Undermine Equilibrium Cooperation

In this section, we show how agents use their communication protocols to extort the principal. The main result of the paper is Proposition 2, which shows that extortion leads cooperation to completely unravel: all agents shirk in every equilibrium.

Cooperation requires that agents communicate among themselves, since without communication an agent would have no way to punish the principal for deviating. We first establish a benchmark result: if agents cannot engage in extortion, then communication can indeed sustain cooperation. To make this point, we consider a **no-extortion game** in which each agent t chooses m_t at the end of period t rather than being committed to μ_t.

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, $(1-\delta)c(e) \leq \delta\bar{\Pi}$. 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 $m_{t'} = C$ for all $t' < t$, then $e_t = e^*$; $s_t = c(e^*)$ if $e_t = e^*$ and $s_t = 0$ otherwise; and $m_t = C$ if neither player deviates and $m_t = D$ otherwise. If $m_{t'} \neq C$ for at least one $t' < t$, then $e_t = s_t = 0$ and $m_t = D$. If $m_{t'} \neq C$ for some $t' < t$, play is as in the one-shot equilibrium and so players cannot profitably deviate. If $m_{t'} = C$ for 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 $(1 - \delta)c(e^*) \leq \delta(e^* - c(e^*))$ since $c(e^*) \leq \delta e^*$. 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. ■

Proposition 1 shows how communication can be used to induce the principal to pay a hard-working agent. 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. The principal would rather pay a transfer than be punished, and each agent would rather exert effort than shirk and forgo the transfer, so this construction is enough to sustain positive effort. If an agent shirks, then his message is independent of the transfer and so the principal pays him nothing. In particular, a shirking agent cannot threaten to send C if the principal pays him and D otherwise. In contrast, we will show that agents make exactly this type of threat in the extortion game.

Next, we present our impossibility result. Consider the mapping from messages to continuation play described in the proof of Proposition 1. In the extortion game, agent t can always choose μ_t so that $m_t = C$ if the principal pays him and $m_t = D$ otherwise. That is, the threat that agent t can make, and hence the compensation he can demand, is essentially independent of his effort. Agents are therefore unwilling to exert effort in equilibrium.

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

Proof: Define $m^{t-1} = (m_0, m_1, ..., m_{t-1})$, 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, with corresponding message $\bar{m}$. Let $\underline{\Pi}$ be the similarly-defined minimum continuation payoff, with corresponding message $\underline{m}$. Agent t is willing to choose $e_t = e^*$ only if $s_t = s^* \geq c(e^*)$; the principal is willing to pay this transfer only if

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

For small $\epsilon > 0$, consider the following deviation by agent t: $e_t = 0$ and

$$\mu_t(s) = \begin{cases} \bar{m} & s = s^* - \epsilon \\ \underline{m} & \text{otherwise.} \end{cases} \tag{2}$$

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

If $s_t > 0$ on the equilibrium path, then 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 was weakly willing to pay s_t when faced with the same punishment, she strictly prefers to pay a smaller amount. An agent can therefore shirk and still guarantee nearly the same transfer as if he had exerted effort. But then each agent shirks, since the transfer he can demand is essentially independent of his effort.

The proof of Proposition 2 relies on the assumption that communication protocols are not publicly observed. Therefore, one might imagine that allowing the principal to communicate

to future agents might improve cooperation by making information about these protocols public. But this turns out not to be true, as online appendix C.1 shows: Proposition 2 continues to hold 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 an agent actually deviated. Online appendix C.2 then shows that cooperation is possible if the principal can *commit* to messages as a function of each agent's effort, provided that she commits to messages *weakly before* each agent chooses his communication protocol. This positive result should be interpreted with skepticism, however, since unlike the agents, the principal sometimes has a strict incentive to deviate from her committed message.[9]

Proposition 2 remains true even if the principal has multiple opportunities to exchange transfers with each agent. In online appendix D.1, we show that our impossibility result continues to hold even if the principal and each agent exchange transfers at both the start and the end of their interactions. Online appendix D.2 then allows each agent to make *ex ante* demands in exchange for refraining from later extortion. Our impossibility result continues to hold even with these *ex ante* demands.

Proposition 2 illustrates a significant roadblock that unions and other organizations must overcome to facilitate cooperation. Even if those organizations coordinate punishments, those punishments face the risk of simply becoming opportunities for the communicating parties to extort their partners. Of course, Proposition 2 is a stark result, and the next two sections explore how cooperation might be sustained even if agents *can* extort the principal. Nevertheless, by identifying how agents can misuse coordinated punishments and how this misuse undermines cooperation, Proposition 2 forms the (pessimistic) foundation for the rest of our analysis.

[9]That is, unlike its role for agents, commitment forces the principal to send messages that are *ex post* suboptimal. Hence, Proposition 8 would not hold in this alternative game: allowing the principal to commit does *not* refine the equilibrium set of the game without commitment.

4 First Remedy: Investigations

This section introduces a public signal of either the effort or the transfer. Depending on the setting, these signals might either arise naturally from an interaction, or be the results of an investigation into each agent's message (by the union, community, association, or online marketplace). We show that these signals limit extortion by tying the threats that an agent can make to his actual interaction with the principal. However, we also show that if these signals are noisy, then extortion continues to constrain equilibrium effort.

4.1 Public Signals of Effort

This subsection studies public signals of each agent's effort. We first prove that if these signals are perfect, then they completely eliminate extortion in equilibrium. Then we present our main result for this section, Proposition 6, which analyzes how even noisy effort signals can support some effort in equilibrium.

Formally, the **extortion game with effort signals** is similar to the extortion game except that after the transfer s_t is paid in each period t, a public signal $z_t \sim G(\cdot|e_t)$, $z_t \in \mathbb{R}$, is realized. Agent t's communication protocol can be any mapping from the transfer *and* this signal to a message, so that (with an abuse of notation) $\mu_t : \mathbb{R}^2 \to M$ and $m_t = \mu_t(s_t, z_t)$.[10] Payoffs are the same as in the extortion game. We focus on two simple signal structures: z_t is **perfect** if $z_t = e_t$; or z_t is **noisy** if $z_t \in \{0, 1\}$ with $\Pr\{z_t = 1|e_t\} = \gamma(e_t)$ for $\gamma(\cdot)$ strictly increasing and twice continuously differentiable.

Future agents can condition their actions on the public signal z_t and thereby stochastically condition their actions on e_t. In the no-extortion game, however, this additional signal would not improve cooperation at all: Proposition 1 holds when there is a signal of effort just as it does without one. The reason is that the proof of Proposition 1 constructs an equilibrium in which agents do not engage in extortion and messages make any deviation by the principal

[10] Allowing μ_t to condition on z_t simplifies our arguments, but the basic intuition would survive if μ_t depended only on s_t.

public knowledge. The effort e^* in a principal-optimal equilibrium is therefore pinned down by the principal's on-path continuation payoff, which means that effort signals are redundant.

In contrast, we show that effort signals substantially increase equilibrium effort in the extortion game by limiting the types of threats that a shirking agent can make and thereby restoring informative communication among the agents. Indeed, perfect effort signals completely eliminate extortion.

Proposition 3 *Define e^* as in Proposition 1. In the extortion game with perfect effort signals, $e_t = e^*$ and $s_t = c(e^*)$ in every $t \geq 0$ of any principal-optimal equilibrium.*

Proof: See appendix A.

To see the proof of Proposition 3, consider a strategy profile with $e_t = e^*$ and $s_t = c(e^*)$ in each period t. To deter deviations, suppose that future agents punish the principal only if an agent reports a deviation *and* the signal reveals that he chose effort e^*. Consequently, if agent t chooses $e_t \neq e^*$, then his message cannot trigger a punishment, so he cannot extort the principal at all. If he chooses $e_t = e^*$, then the principal's punishment can be made just severe enough that she is willing to pay $s_t = c(e^*)$ but no more. The resulting equilibrium implements effort e^* and gives all of the resulting surplus to the principal.

Proposition 3 illustrates how effort signals can restore cooperation under the extreme assumption that signals are perfect. If the signal is noisy, then a shirking agent might go undetected by future agents. However, the amount that an agent can extort can still be tied to his effort *in expectation*, which means that positive effort can be sustained in equilibrium. We also show that agents earn rent that is increasing in their equilibrium effort.

Proposition 4 *Suppose $\gamma(e_t)$ is weakly concave in the extortion game with noisy effort signals. Then:*

1. *There exists a strictly increasing function $\bar{u} : \mathbb{R}_+ \to \mathbb{R}_+$ with $\bar{u}(0) = 0$ such that if agent t chooses e_t on the equilibrium path, then his payoff is at least $\bar{u}(e_t)$.*

2. *There exists an equilibrium in which $e_t = e^*$ in each $t \geq 0$ if and only if*

$$c'(e^*) \leq \gamma'(e^*)\frac{\delta}{1-\delta}(e^* - c(e^*) - \bar{u}(e^*)). \tag{3}$$

3. *There exists $\bar{\delta} < 1$ such that if $\delta \geq \bar{\delta}$, the effort in each $t \geq 0$ of any principal-optimal equilibrium equals*

$$e^* \in \arg\max_e \{e - c(e) - \bar{u}(e)\} < e^{FB}.$$

Proof: See appendix A.

The proof of Proposition 4 exhibits a close connection between equilibrium play in the extortion game with noisy effort signals and a static contracting problem with limited liability. Let $\bar{\Pi}(z_t)$ and $\underline{\Pi}(z_t)$ be the largest and smallest principal continuation payoffs following signal realization z_t in period t. Define

$$\hat{s}(e_t) \equiv \frac{\delta}{1-\delta}\mathbb{E}\left[\bar{\Pi}(z_t) - \underline{\Pi}(z_t)|e_t\right] \geq 0.$$

The principal is willing to pay s_t only if $s_t \leq \hat{s}(e_t)$, since otherwise she would rather renege and earn continuation payoff $\mathbb{E}\left[\underline{\Pi}(z_t)|e_t\right]$. However, agent t can use a communication protocol similar to (2) in order to extort any $s_t < \hat{s}(e_t)$. Agent t therefore chooses e_t to maximize $\hat{s}(e_t) - c(e_t)$. Letting $R(z_t) \equiv \bar{\Pi}(z_t) - \underline{\Pi}(z_t)$, agent t chooses e_t in equilibrium only if

$$e_t \in \arg\max_e \left\{\mathbb{E}\left[R(z)|e\right] - c(e)\right\}. \tag{4}$$

Since $R(\cdot) \geq 0$, this incentive constraint is similar to that of a static contracting problem with limited liability. It follows that agent t earns a rent $\bar{u}(e_t)$ that is increasing in his equilibrium effort e_t.

As in the static contracting problem, agent t's effort incentives are strongest if $R(0) = 0$. If $e_t = e^*$ in each period t, then $R(1) \leq e^* - c(e^*) - \bar{u}(e^*)$. Setting $R(0) = 0$ and $R(1) =$

$e^* - c(e^*) - \bar{u}(e^*)$ in the first-order condition of (4) yields (3). Since $R(\cdot)$ depends only on the difference $\bar{\Pi}(\cdot) - \underline{\Pi}(\cdot)$, we can increase $\bar{\Pi}(\cdot)$ without changing either agent t's incentives or his payoff. Therefore, principal-optimal equilibria always entail principal-optimal continuation play and hence stationary effort on the equilibrium path. In those equilibria, effort is strictly lower than first-best because a slight decrease in effort at e^{FB} entails a second-order loss in total surplus but a first-order decrease in agent t's rent.

The fact that agents earn rent means that the principal has less to lose from reneging on any given agent, since she earns only a fraction of the value created in future interactions. She is therefore less willing to pay agents, who in turn are less willing to exert effort. Consequently, $\bar{u}(e^*)$ enters negatively on the right-hand side of (3), leading to a lower maximum effort than in the no-extortion game. This distortion, which resembles a distortion in Proposition 5 of Levin (2002), implies that an agent's rent-seeking behavior imposes a negative externality on *other* agents by making the principal's promises to those agents less credible.

In practice, the players might have some sway over the signal distribution, as, for instance, when a union decides how to investigate grievances. While both agents and the principal prefer some kind of investigation to none, Propositions 3 and 4 suggest that these parties have different preferences over the precision of that investigation. In a principal-optimal equilibrium, both total surplus and the principal's payoff are maximized if signals are perfect. The agents, however, earn rent only if the signal is noisy, so they might collectively prefer to implement less precise investigations.

4.2 Public Signals of Transfers

We now turn to public signals of the transfer. As in section 4.1, we first show how perfect signals can eliminate extortion, then we analyze equilibrium effort if signals are noisy.

The **extortion game with transfer signals** is identical to the extortion game except that a public signal $x_t \sim F(\cdot|s_t)$, $x_t \in \mathbb{R}$, is realized after s_t in each t and observed by all players. Agent t's communication protocol 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 focus on either **perfect transfer signals,** $x_t = s_t$; or **noisy transfer signals,** $x_t \in \{0, 1\}$ with $\Pr\{x_t = 1|s_t\} = \phi(s_t)$ for some strictly increasing and twice continuously differentiable $\phi(\cdot)$.

As with a signal of e_t, a public signal of s_t does not change the set of equilibrium outcomes at all in the no-extortion game. However, once we allow agents to extort the principal, this signal can lead to more cooperation in equilibrium. Indeed, we show that a perfect transfer signal is enough to recover the equilibrium effort level from Proposition 1.

Proposition 5 *Define e^* as in Proposition 1. In the extortion game with perfect transfer signals, $e_t = e^*$ and $s_t = c(e^*)$ in every $t \geq 0$ of any principal-optimal equilibrium.*

Proof: See appendix A.

The statement of Proposition 5 is nearly identical to that of Proposition 3, but the underlying intuition is quite different. Since x_t does not directly reveal anything about agent t's actions, it cannot be used to tie the amount he can extort to his effort. We show that x_t can instead be used to make the principal indifferent between paying an agent or not, so that she is both willing to pay $c(e^*)$ if $e_t = e^*$ and willing to pay nothing if $e_t \neq e^*$. In particular, suppose future agents punish the principal unless $m_t = C$ *and* $x_t = c(e^*)$. If agent t shirks, then he can try to extort $s_t = c(e^*)$ but no less, since any smaller transfer triggers punishment regardless of the message. If the punishment is such that the principal is exactly indifferent between paying $s_t = c(e^*)$ and $s_t = 0$, then she is willing to pay $s_t = c(e^*)$ if $e_t = e^*$, and $s_t = 0$ otherwise. But then agent t cannot earn more than zero by deviating, so he is willing to choose $e_t = e^*$.

This equilibrium construction hinges on two features: (i) the principal is indifferent between paying $s_t = c(e^*)$ and $s_t = 0$, and (ii) she cannot be forced to pay any other amount. These two conditions are relatively easy to satisfy with perfect signals. With noisy signals, however, they dramatically restrict the efforts and transfers that can be supported in equilibrium. Our next result formalizes these restrictions as a set of necessary conditions

for equilibrium.

Lemma 1 *Suppose x_t is a noisy signal of s_t. If $e_t > 0$ on-path in an equilibrium, then there must exist $s^* > 0$ and $\hat{s} \in [0, s^*)$ such that (i) $c(e_t) \leq s^* - \hat{s}$, (ii) $\phi''(s^*) \leq 0$, and (iii)*

$$\phi'(s^*) = \frac{\phi(s^*) - \phi(\hat{s})}{s^* - \hat{s}}. \tag{5}$$

Proof: See appendix A.

Suppose that $\bar{\Pi}(x_t)$ and $\underline{\Pi}(x_t)$ are the principal's largest and smallest equilibrium continuation payoffs following signal x_t. Define

$$\pi^D = \max_{s_t \geq 0} \{-(1-\delta)s_t + \delta \mathbb{E}[\underline{\Pi}(x_t)|s_t]\}$$

as the principal's *equilibrium* min-max payoff from period t onwards, *including* her disutility from paying s_t. The principal is willing to pay s_t only if

$$-(1-\delta)s_t + \delta \mathbb{E}\left[\bar{\Pi}(x_t)|s_t\right] \geq \pi^D. \tag{6}$$

If agent t shirks, then he can extort any transfer s_t that strictly satisfies (6). Let $\hat{s}$ be the supremum of the set of transfers that can be extorted.

Since $\mathbb{E}\left[\bar{\Pi}(x_t)|s_t\right]$ is continuous in s_t, agent t's supremum payoff from a deviation equals $\hat{s}$, which moreover must satisfy (6) with equality. Agent t is willing to choose $e_t > 0$ in equilibrium only if doing so and receiving s^* is better than shirking and extorting $\hat{s}$, that is, if $s^* - c(e_t) \geq \hat{s}$. But (6) must hold with equality at $s_t = s^*$, since otherwise $\hat{s} \geq s^*$. The left-hand side of (6) must also be concave and tangent to the right-hand side at $s_t = s^*$, since otherwise (6) would hold strictly on at least one side of s^*, and thus again $\hat{s} \geq s^*$. Given that (6) holds with equality at both $\hat{s}$ and s^*, the *average* slope of $\mathbb{E}[\bar{\Pi}(x_t)|s_t]$ between $\hat{s}$ and s^* equals that of the line tangent to $\mathbb{E}\left[\bar{\Pi}(x_t)|s_t = s^*\right]$. Equation (5) represents this equality for this signal structure.

Lemma 1 implies that inducing cooperation in the extortion game with noisy transfer signals requires $\phi(\cdot)$ to satisfy several restrictive conditions. Our next result shows that these conditions are impossible to satisfy if $\phi(\cdot)$ is concave, which means that positive effort is possible only if $\phi(\cdot)$ has both convex and concave regions. We demonstrate that positive effort is sometimes possible in that case.

Proposition 6 *Suppose x_t is a noisy signal of s_t. If $\phi(\cdot)$ is strictly concave on $\mathbb{R}_+$, then $e_t = 0$ in each $t \geq 0$ of every equilibrium.*

Suppose that on $\mathbb{R}_+$, $\phi(\cdot)$ is first strictly convex and then strictly concave. Let $e^* > 0$, $s^* > 0$, and $\hat{s} \in [0, s^*)$ satisfy the conditions of Lemma 1, with $e^* - c(e^*) - \hat{s} > 0$. Then there exists $\bar{\delta} \in [0, 1)$ such that for any $\delta \geq \bar{\delta}$, there exists an equilibrium in which $e_0 = e^*$. Any such equilibrium is nonstationary.

Proof: See appendix A.

To prove the first part of Proposition 6, it is enough to note that the slope of any line that connects two points on a strictly concave $\phi(\cdot)$ cannot coincide with the tangent line at either of those points, which means that (5) cannot be satisfied. The second part of Proposition 6 investigates the simplest possible case that might satisfy the conditions of Lemma 1: $\phi(\cdot)$ is first convex and then concave. In that case, if there exist e^*, s^*, and $\hat{s}$ that satisfy these conditions, then we can work backwards from the proof of Lemma 1 to construct an equilibrium that implements effort e^* in the first period. This construction requires the principal to be punished whenever $x_t = 0$, since otherwise $\mathbb{E}[\bar{\Pi}(x_t)|s_t]$ would not vary with s_t. The principal must therefore be periodically punished in any equilibrium with positive effort.

Together, sections 4.1 and 4.2 highlight two conditions that must hold for extortion to lead to zero equilibrium effort. First, agents must be able to shirk without being detected by other agents. Second, after an agent shirks, he must be able to give the principal a *strict* incentive to pay nearly the same transfer that she would pay a hard-working agent.

Effort signals restore cooperation by tailoring an agent's ability to extort to his effort, which violates the first condition. Transfer signals restrict agents' threats as a function of the transfer, which violates the second condition. However, while either type of investigation can improve cooperation, each comes with its own weaknesses. Effort signals typically lead to agents earning rent in equilibrium, which undermines the principal's willingness to pay large transfers, while transfer signals require an equilibrium construction that both is fragile and typically entails on-path punishments.

5 Second Remedy: Bilateral 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 future *bilateral* interactions between the principal and each agent can be used to support cooperation. As is familiar from the literature on repeated games, these bilateral interactions can be used to punish an agent for shirking or the principal for reneging on a hard-working agent, leading to positive equilibrium effort. We now emphasize a third effect that is specific to our setting: bilateral relationships can be used to punish the principal for acquiescing to an agent's threats, which makes extortion less tempting to each agent. Therefore, bilateral relationships enable coordinated punishments.

Consider the **extortion game with bilateral relationships**, which is identical to the extortion game except 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 agent. 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 $e_t - s_t + v_t$ and $s_t - c(e_t) + v_t$, respectively.

The outcomes of the coordination game are not observed by any future agents and so

cannot affect the principal's continuation value. In equilibrium, v_t must therefore correspond to a Nash equilibrium of the coordination game in each $t \geq 0$. Define $v_t = H$ and $v_t = L$ as the largest and smallest such Nash equilibrium payoffs, respectively. While our result can be readily extended for general, possibly asymmetric coordination games, the following simple example suffices:

$$\begin{array}{ccc} & h & l \\ h & (H,H) & (L,L) \\ l & (L,L) & (L,L) \end{array}.$$

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

Proposition 7 *In the extortion game with bilateral relationships, $c(e_t) \leq 3(H-L)$ in every $t \geq 0$ of any equilibrium. If e^* is the minimum of e^{FB} and the solution to $c(e^*) = 3(H-L)$, 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 effort constraint $c(e_t) \leq 3(H - L)$ reflects the fact that bilateral relationships can encourage equilibrium effort via three different channels. Two of these channels are familiar: agent t can be punished by $(H - L)$ if he fails to exert the equilibrium effort level, and the principal can be punished by $(H - L)$ if she fails to reward an agent who exerts effort. We show that the coordination game can also be used to reward the principal for refusing to pay an agent who shirks, so that the principal is willing to pay $(H - L)$ more to an agent who exerts effort relative to one who tries to extort her. Consequently, agents' messages can influence equilibrium continuation play without opening the door to extortion. The proof of Proposition 7 simply tracks the histories associated with each of these three channels and

arranges transfers so that the appropriate players are rewarded or punished at each of those histories. Note that, unlike in the rest of the paper, this proof uses the fact that agents can pay the principal.

As in the proof of Proposition 2, let us define $\bar{\Pi}$ and $\underline{\Pi}$ as the principal's largest and smallest equilibrium continuation payoffs, respectively, in some period t, in which case agent t's messages can punish the principal by no more than $\delta(\bar{\Pi}-\underline{\Pi})$. Agent t can therefore extort no more than $\delta(\bar{\Pi}-\underline{\Pi})-(1-\delta)(H-L)$ if he shirks, since the principal loses $(H-L)$ from her bilateral relationship with agent t if she gives in to extortion. That is, agent t cannot extort the principal at all as long as

$$\frac{\delta}{1-\delta}\left(\bar{\Pi}-\underline{\Pi}\right) \leq H-L. \tag{7}$$

It is in this sense that bilateral relationships enable coordinated punishments: a larger difference $(H-L)$ implies that the coordinated punishment $\bar{\Pi}-\underline{\Pi}$ can be more severe before it leads to extortionary threats. On the other hand, if (7) is violated, then increasing $\bar{\Pi}-\underline{\Pi}$ increases both agent t's on-path payment and the amount he can extort, and so could not increase agent t's equilibrium effort.

The coordination game is an abstract way to reflect the idea that the principal has more than one interaction with each agent. In reality, these bilateral relationships are typically long-lived: 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 can interpret the coordination-game payoff v_t as a simple representation of the continuation payoff from these future interactions. In appendix B, we confirm this interpretation by studying a setting with truly long-run agents who interact repeatedly with the principal. While the resulting analysis is more involved, it remains true that bilateral relationships facilitate coordinated punishments.

6 Interpreting the Communication Protocol

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

Without the communication protocol 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 communication protocol 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 8 *For any equilibrium of the extortion game or of the extortion game with effort signals, transfer signals, or bilateral 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: See appendix A.

Since agents are indifferent among messages, they are always willing to follow through on their communication protocols. 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 communication protocols, 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 communication protocol is enough to lead to the complete collapse of effort in equilibrium.

Proposition 9 *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: See appendix A.

Even a small preference for following the communication protocol 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 communication protocols if $\epsilon > 0$ is small.

Proposition 9 assumes that agents prefer to "keep their word" by acting according to their communication protocol. 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 benchmark 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 8 and 9 suggest two reasons why commitment is a relatively mild assumption in our setting: first, it simply refines the set of equilibria, and second, many of our results hold even if we replaced commitment with a mild preference for following the communication protocol. Fundamentally, however, the applications in our introduction are what motivate us to study the extortion game. The threat of extortion features prominently in each of these applications, and studying extortion requires a setting in which agents can make action-contingent threats even after they deviate. The communication protocol, or something like it, is therefore necessary to study this kind of extortion and identify new ways to encourage cooperation.

7 Conclusion

In many settings, businesses and individuals cooperate with one another because they expect partners to spread word of any misbehavior. This paper studies an underexplored obstacle to using communication as a way to coordinate punishments: agents may misuse messages intended to report deviations in order to extort the principal instead. 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 extortionary threats are unlikely to be detected. Second, coordinated punishments are valuable when bilateral relationships are relatively weak, which means that the extorted party has little direct recourse to punish the extorter. While extortion poses a significant challenge to cooperation, our remedies highlight how organizations could mitigate its negative consequences and restore the use of coordinated punishments.

Our analysis suggests two natural next steps. First, we could delve further into the assumption that agents commit to their communication protocols. We have argued that this assumption is both reasonable and (in a sense) necessary to study extortion, which requires agents to make off-path, action-contingent threats. We could therefore ask: under what

circumstances does an agent have the incentive to develop a reputation for following through on these threats? If other agents know about that reputation, then the extorting agent's message might simply be ignored in equilibrium. However, if an agent's reputation is known to the principal but not to other agents, then he has every incentive to build a reputation with the principal for extortion.[11] Exploring agents' incentives to build reputations for extortion would complement the analysis in this paper and point to other ways that the principal might mitigate the negative consequences of extortion.

Second, the agents rely on communication to coordinate punishments but the principal does not. One could study extortionary threats in settings with more symmetric interactions, as in, for example, a network of relationships (e.g., Ali and Miller (2016)). In contrast to our setting, in a more symmetric transaction *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 these questions.

[11] In a sense, the communication protocol allows each agent to act as a "Stackleberg type" in his private interaction with the principal.

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.

Awaya, Y. and V. Krishna (2018). Communication and cooperation in repeated games. Working Paper.

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.

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.

Compte, O. (1998). Communication in repeated games with imperfect private monitoring. *Econometrica 66*(3), 597–626.

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

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.

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.

Guo, Y. and E. Shmaya (2018). Costly miscalibration. 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.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Kandori, M. and H. Matsushima (1998). Private observation, communication, and collusion. *Econometrica 66*(3), 627–652.

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.

Lipnowski, E., D. Ravid, and D. Shishkin (2018). Persuasion via weak institutions. Working Paper.

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

MacLeod, B. and J. Malcomson (1989). Implicit contracts, incentive compatibility, and involuntary unemployement. *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.

Meisner, J. and B. Ruthhart (2017). Teamsters boss indicted on charges of extorting $100,000 from business. *Chicago Tribune*.

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.

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

Rayo, L. and I. Segal (2010). Optimal information disclosure. *Journal of Political Economy 118*(5), 949–987.

Watson, J. (2016). Perfect bayesian equilibrium: General definitions and illustrations.

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.

A Omitted Proofs

A.1 Proof of Proposition 3

The first paragraph of the proof of Proposition 1 applies to the extortion game with perfect effort signals as well. Therefore, the equilibrium surplus is at most $e^* - c(e^*)$. Moreover, if $e^* < e^{FB}$, then the highest equilibrium effort is at most e^*.

We now construct an equilibrium in which $e_t = e^*$ in every period. The equilibrium starts in the cooperative phase. In this phase, each agent t chooses $e_t = e^*$ and

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

The principal pays $s_t = c(e^*)$ if agent t does not deviate, $s_t = 0$ if $e_t \neq e^*$, and best-responds to μ_t if $e_t = e^*$ but $\mu_t(\cdot)$ is not the above protocol. If either $m_t \neq C$ or $z_t \neq e^*$, play switches to the punishment phase with probability $\alpha \in [0, 1]$ that satisfies

$$c(e^*) = \frac{\delta}{1-\delta} \alpha \left(e^* - c(e^*)\right). \tag{8}$$

In the punishment phase, each agent t chooses $e_t = 0$ and the principal chooses $s_t = 0$. If $m_t = C$ and $z_t = e^*$, play stays in the cooperative phase.

If agent t chooses $e_t = e^*$, then (8) implies that the principal is willing to pay no more than $c(e^*)$; for any other effort, the principal pays $s_t = 0$. Agent t therefore has no profitable deviation. The equation (8) implies that the principal is indifferent between $s_t = c(e^*)$ and $s_t = 0$ if $e_t = e^*$. For any $e_t \neq e^*$, the principal's continuation payoff is independent of s_t and so $s_t = 0$ is a best response. Therefore, the principal has no profitable deviation.

This equilibrium construction maximizes total surplus and gives all of that surplus to the principal. It is therefore principal-optimal, and any principal-optimal equilibrium must entail the same efforts and transfers. ■

A.2 Proof of Proposition 4

Statement 1

Suppose $e_t = e^*$ in period t of an equilibrium. Let $\bar{\Pi}(z)$ and $\underline{\Pi}(z)$ be the principal's largest and smallest continuation payoffs following signal realization z, and let $\bar{m}(z)$ and $\underline{m}(z)$ be the corresponding messages. Agent t can always choose e_t and

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

The principal's unique best response is to pay $\hat{s}$ so long as

$$\hat{s} < \hat{s}(e_t) \equiv \frac{\delta}{1-\delta}\left(\gamma(e_t)(\bar{\Pi}(1) - \underline{\Pi}(1)) + (1-\gamma(e_t))(\bar{\Pi}(0) - \underline{\Pi}(0)\right). \tag{9}$$

Clearly, the principal is willing to pay no more than $\hat{s}(e_t)$ following effort e_t. Therefore, $e^* \in \arg\max_e \{\hat{s}(e) - c(e)\}$.

Since $\gamma(\cdot)$ is concave, so is $\hat{s}(\cdot)$. Since $c'(0) = 0$, e^* is characterized by the first-order condition

$$c'(e^*) = \hat{s}'(e^*) = \frac{\delta}{1-\delta}\gamma'(e^*)\left(\bar{\Pi}(1) - \underline{\Pi}(1) - \left(\bar{\Pi}(0) - \underline{\Pi}(0)\right)\right). \tag{10}$$

Since $\bar{\Pi}(0) \geq \underline{\Pi}(0)$,

$$\hat{s}(e^*) - c(e^*) \geq \gamma(e^*)\frac{c'(e^*)}{\gamma'(e^*)} - c(e^*) \equiv \bar{u}(e^*),$$

where this inequality follows by setting $\bar{\Pi}(0) = \underline{\Pi}(0)$ in (9) and then substituting $(\delta/(1-\delta))\left(\bar{\Pi}(1) - \underline{\Pi}(1) - \left(\bar{\Pi}(0) - \underline{\Pi}(0)\right)\right)$ from (10).

The fact that $\bar{u}(0) = 0$ is immediate from $c(0) = c'(0) = 0$. Since $c(\cdot)$ is strictly increasing

and strictly convex and $\gamma(\cdot)$ is strictly increasing and weakly concave,

$$\bar{u}'(e_t) = -\frac{\gamma''(e_t)\gamma(e_t)}{\gamma'(e_t)}\frac{c'(e_t)}{\gamma'(e_t)} + \frac{\gamma(e_t)}{\gamma'(e_t)}c''(e_t) > 0,$$

so $\bar{u}(\cdot)$ is strictly increasing. $\square$

Statement 2

In an equilibrium with $e_t = e^*$ in each period t, $\bar{\Pi}(z) - \underline{\Pi}(z) \leq e^* - c(e^*) - \bar{u}(e^*)$ for each $z \in \{0, 1\}$. Therefore, (10) requires that

$$c'(e^*) \leq \frac{\delta}{1-\delta}\gamma'(e^*)\left(e^* - c(e^*) - \bar{u}(e^*)\right).$$

So (3) is necessary for $e_t = e^*$ in each period.

Now, suppose that (3) holds. Consider the following strategy profile. In the cooperative phase, each agent t chooses $e_t = e^*$ and

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

The principal pays $s_t = c(e^*) + \bar{u}(e^*)$ if agent t does not deviate, and otherwise chooses the smallest s_t that is a best response to $\mu_t(s, z)$ given e_t. Play stays in the cooperative phase until $m_t = D$ and $z_t = 1$, at which point it switches to the punishment phase with probability $\alpha \in [0, 1]$. The punishment phase is absorbing and entails $e_t = s_t = 0$ in each period.

Suppose that $\alpha \in [0, 1]$ satisfies

$$c'(e^*) = \gamma'(e^*)\frac{\delta}{1-\delta}\alpha\left(e^* - c(e^*) - \bar{u}(e^*)\right).$$

Following effort e_t, agent t can earn no more than

$$\frac{\delta}{1-\delta}\alpha\gamma(e_t)\left(e^* - c(e^*) - \bar{u}(e^*)\right) = \gamma(e_t)\frac{c'(e^*)}{\gamma'(e^*)} - c(e_t).$$

This expression is concave in e_t and maximized at $e_t = e^*$, so agent t has no profitable deviation from his strategy. If agent t does not deviate, then the principal pays $c(e^*) + \bar{u}(e^*)$ so long as

$$c(e^*) + \bar{u}(e^*) \leq \gamma(e^*)\frac{\delta}{1-\delta}\alpha(e^* - c(e^*) - \bar{u}(e^*)),$$

which holds by definition of $\bar{u}(e^*)$ and α. So the principal has no profitable deviation. We conclude that an equilibrium with $e_t = e^*$ in each $t \geq 0$ exists whenever (3) holds. □

Statement 3

Let Π^* be the principal's payoff in the principal-optimal equilibrium. Then $\Pi^* \leq (1-\delta)\arg\max_{e\geq 0}\{e - c(e) - \bar{u}(e)\} + \delta\Pi^*$ because agent t earns no less than $\bar{u}(e)$ if $e_t = e$. So $\Pi^* \leq \arg\max_{e\geq 0}\{e - c(e) - \bar{u}(e)\}$. Let e^* maximize $e - c(e) - \bar{u}(e)$. Statement 2 says that there exists $\bar{\delta} < 1$ such that for $\delta \geq \bar{\delta}$, there exists an equilibirum with $e_t = e^*$ in each $t \geq 0$. The proof of this statement constructs an equilibrium of this kind in which each agent earns $\bar{u}(e^*)$. The principal therefore earns $e^* - c(e^*) - \bar{u}(e^*) = \Pi^*$. Moreover, any equilibrium that attains this payoff must have $e_t = e^*$ (with probability 1) in every period on the equilibrium path. ■

A.3 Proof of Proposition 5

The first paragraph of the proof for Proposition 1 applies to the extortion game with perfect transfer signals as well. Therefore, the equilibrium surplus is at most $e^* - c(e^*)$. Moreover, if $e^* < e^{FB}$, then the highest equilibrium effort is at most e^*.

We now construct an equilibrium in which $e_t = e^*$ and $s_t = c(e^*)$ in every $t \geq 0$. Play starts in the cooperative phase: in each $t \geq 0$, agent t chooses e^* and constant $\mu^*(\cdot)$. If

$e_t = e^*$, then $s_t = c(e^*)$; otherwise, $s_t = 0$. If $x_t \neq c(e^*)$, then play switches to the punishment phase with probability α and otherwise remains in the cooperative phase. The punishment phase is absorbing and features $e_t = s_t = 0$ in each t.

Fix $\alpha \in [0, 1]$ so that

$$c(e^*) = \frac{\delta}{1-\delta}\alpha(e^* - c(e^*)). \tag{11}$$

Agents have no profitable deviation because they earn no more than 0 from deviating. By (11), the principal is willing to pay $s_t = c(e^*)$ in the cooperative phase. The principal is also willing to pay $s_t = 0$ because (11) holds with equality. Therefore, the principal has no profitable deviation either. So this strategy profile is an equilibrium with the desired properties. ■

A.4 Proof of Lemma 1

Consider an equilibrium with $e_t = e^* > 0$. Define $\bar{\Pi}(x)$, $\underline{\Pi}(x)$ as the largest and smallest principal continuation payoffs following x, with corresponding messages $\bar{m}(x)$ and $\underline{m}(x)$. Define

$$\pi^D = \max_{s \geq 0} \{-(1-\delta)s + \delta\left(\phi(s)\underline{\Pi}(1) + (1-\phi(s))\underline{\Pi}(0)\right)\} \tag{12}$$

as the principal's min-max payoff as a function of the period-t transfer s. Let $s_A \geq 0$ be the *smallest* maximizer of (12). Define

$$s_B \equiv \sup\left\{s \geq 0 | -(1-\delta)s + \delta\left(\phi(s)\bar{\Pi}(1) + (1-\phi(s))\bar{\Pi}(0)\right) > \pi^D\right\}$$

if this supremum exists. Note that $s_B > 0$ whenever it exists.

If s_B exists, then agent t can always deviate by choosing $e_t = 0$ and

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

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

Similarly, agent t can always deviate by choosing $e_t = 0$ and

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

The principal may have multiple best responses to this deviation. However, all best responses entail $s_t \geq s_A$, since for any $s < s_A$,

$$-(1-\delta)s_A + \delta\mathbb{E}[\bar{\Pi}(x)|s_A] \geq \pi^D > -(1-\delta)s + \delta\mathbb{E}[\underline{\Pi}(x)|s],$$

where the first inequality holds because $\bar{\Pi}(\cdot) \geq \underline{\Pi}(\cdot)$ and the second inequality holds because s_A is the smallest maximizer of (12).

Now, define $\hat{s} = s_A$ if s_B does not exist and $\hat{s} = \max\{s_A, s_B\}$ otherwise. Agent t can guarantee a payoff arbitrarily close to $\hat{s}$ if he deviates, so he chooses $e_t = e^*$ only if $s^* - c(e^*) \geq \hat{s}$. Moreover,

$$\begin{gathered} -(1-\delta)s^* + \delta\left(\phi(s^*)\bar{\Pi}(1) + (1-\phi(s^*))\bar{\Pi}(0)\right) = \\ \pi^D = \\ -(1-\delta)\hat{s} + \delta\left(\phi(\hat{s})\bar{\Pi}(1) + (1-\phi(\hat{s})\bar{\Pi}(0)\right) \end{gathered} \tag{13}$$

where the first equality follows because $s^* > \hat{s}$ and the principal must be willing to pay s^*, while the second equality follows by continuity of $\phi(\cdot)$ and the definition of $\hat{s}$.

Rearranging, we have

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

Given (13),

$$-(1-\delta)s + \delta\left(\phi(s)\bar{\Pi}(1) + (1-\phi(s))\bar{\Pi}(0)\right)$$

must attain a local maximum at $s = s^*$, since otherwise it would be strictly larger than π^D on at least one side of $s = s^*$ and so $\hat{s} \geq s^*$. Equilibrium therefore requires

$$\phi'(s^*)\left(\bar{\Pi}(1) - \bar{\Pi}(0)\right) = \frac{1-\delta}{\delta} \tag{14}$$

and $\phi''(s^*) \leq 0$, our second necessary condition. Together, (14) and (13) imply (5), our final necessary condition. ■

A.5 Proof of Proposition 6

Suppose that $\phi(\cdot)$ is strictly concave on $\mathbb{R}_+$. Then for any $s^* > \hat{s} \geq 0$,

$$\phi(s^*) - \phi(\hat{s}) = \int_{\hat{s}}^{s^*} \phi'(s)ds < \int_{\hat{s}}^{s^*} \phi'(s^*)ds = \phi'(s^*)(s^* - \hat{s}),$$

so (5) cannot hold. Lemma 1 implies that $e_t = 0$ in every t of any equilibrium.

Now, suppose that $\phi(\cdot)$ is first strictly convex and then strictly concave on $\mathbb{R}_+$. Consider the following strategy profile. Play begins in the cooperative phase. In each period t of the cooperative phase: agent t chooses $e_t = e^*$ and

$$\mu_t(s,x) = \begin{cases} C & \text{if } s_t = c(e^*) + \hat{s} \\ D & \text{otherwise.} \end{cases}$$

If agent t has not deviated, the principal pays $s_t = c(e^*) + \hat{s}$. Otherwise, the principal chooses the minimum s_t that is a best response to μ_t. If $x_t = 1$ and $m_t = C$, then stay in the cooperative phase; if $x_t = 0$ and $m_t = C$, then transition to the punishment phase with probability α_C; if $m_t = D$, then transition to the punishment phase with probability α_D. In the punishment phase, $e_{t'} = s_{t'} = 0$ for all $t' \geq t$.

Define $s^* = c(e^*) + \hat{s}$. In the cooperative phase, the principal's expected continuation payoff at the start of each period, Π^*, satisfies

$$\frac{\delta}{1-\delta}\Pi^* = \frac{\delta(e^* - c(e^*) - \hat{s})}{1 - \delta\left(\phi(s^*) + (1-\phi(s^*))(1-\alpha_C)\right)}.$$

Her continuation payoff after $m_t = D$ equals $(1-\alpha_D)\Pi^*$.

Define α_C so that

$$\frac{1}{\phi'(s^*)} = \alpha_C \frac{\delta}{1-\delta}\Pi^*. \tag{15}$$

Define α_D so that

$$s^* = \frac{\delta}{1-\delta}\left(\phi(s^*)\Pi^* + (1-\phi(s^*))(1-\alpha_C)\Pi^* - (1-\alpha_D)\Pi^*\right).$$

For δ sufficiently close to 1, both of these conditions can be satisfied with $\alpha_C, \alpha_D \in [0,1]$.

We show that this strategy profile is an equilibrium. If agent t has not deviated, the principal strictly prefers $s_t \in \{0, s^*\}$ to any other s_t, and α_D is such that the principal is indifferent between $s_t = 0$ and $s_t = s^*$. So the principal has no profitable deviation from $s_t = s^*$.

If agent t has deviated, then we show that the principal is willing to pay no more than $\hat{s}$. Define

$$\Pi_E(s) = \frac{\delta}{1-\delta}\left(\phi(s)\Pi^* + (1-\phi(s))(1-\alpha_C)\Pi^*\right)$$

as the principal's continuation payoff if $m = C$, and note that

$$\Pi_E''(s) = \frac{\delta}{1-\delta}\phi''(s)\alpha_C\Pi^*.$$

Define

$$\psi(s) = \frac{\delta}{1-\delta}\left(\phi(s^*)\Pi^* + (1-\phi(s^*))(1-\alpha_C)\Pi^*\right) + s - s^*$$

as the 45-degree line that coincides with $\Pi_E(s)$ at s^*.

Condition (15) implies that $\Pi_E(s)$ is tangent to $\psi(s)$ at s^*; since $\phi''(s^*) < 0$, we conclude that s^* is the only point in the concave region of $\phi(\cdot)$ that satisfies $\Pi_E(s) = \psi(s)$. In the strictly convex region of $\phi(\cdot)$, $\Pi_E(s)$ is strictly convex, and by the above argument, $\Pi_E(s)$ lies below $\phi(s)$ at the right endpoint of this convex region. Consequently, $\Pi_E(s)$ crosses $\psi(s)$ at most once in the convex region, and this crossing must be from above. Moreover, $\Pi_E(\hat{s}) = \psi(\hat{s})$ is the unique crossing in the convex region, since

$$\begin{aligned} \psi(\hat{s}) - \Pi_E(\hat{s}) &= \tfrac{\delta}{1-\delta}\left(\phi(s^*) - \phi(\hat{s})\right)\alpha_C\Pi^* + \hat{s} - s^* \\ &= \tfrac{\phi(s^*)-\phi(\hat{s})}{\phi'(s^*)} + \hat{s} - s^* \\ &= 0, \end{aligned}$$

where the first equality follows from the definition of α_C and the second equality follows from (5).

We claim that agent t can deviate to $e_t = 0$ and choose μ_t so that the principal's unique best response is $s_t = s$ exactly when $\psi(s) < \Pi_E(s)$. To see this, note that

$$\psi(s) = s^* + (1-\alpha_D)\Pi^* + s - s^* = s + (1-\alpha_D)\frac{\delta}{1-\delta}\Pi^*$$

by definition of α_D. Consequently, $\psi(s) < \Pi_E(s)$ exactly when

$$s < \frac{\delta}{1-\delta}\left(\phi(s)\Pi^* + (1-\phi(s))(1-\alpha_C)\Pi^* - (1-\alpha_D)\Pi^*\right).$$

If this inequality is satisfied, then agent t can force the principal to pay $s_t = s$. Conversely, if this inequality does not hold, then the principal cannot be induced to pay s, since she earns no less than $\delta(1-\alpha_D)\Pi^*$ from paying 0 and no more than $-(1-\delta)s + \delta\Pi_E(s)$ from paying s.

We have already shown that $\psi(s) \geq \Pi_E(s)$ for all $s \geq \hat{s}$, which means agent t earns no more than $\hat{s}$ if he chooses $e_t \neq e^*$. We conclude that he has no profitable deviation from e^*, since $c(e^*) \leq s^* - \hat{s}$. Consequently, the proposed strategy profile is an equilibrium for δ

sufficiently close to 1. ■

A.6 Proof of Proposition 7

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 $e_t = 0$ 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

$$\hat{s} < L - H + \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}). \tag{16}$$

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

$$s^* \leq H - L + \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}). \tag{17}$$

Agent t is willing to choose $e_t = e^*$ only if $s^* - c(e^*) + (H - L) \geq \hat{s}$ for *any* $\hat{s}$ satisfying (16). Given the bound (17) on s^*, we conclude that $e_t = e^*$ in equilibrium only if $3(H-L) \geq c(e^*)$.

Each agent must earn at least L, so the principal's equilibrium payoff cannot exceed $e^* - c(e^*) + 2H - L$, where $e^* = e^{FB}$ if $c(e^{FB}) \leq 3(H - L)$ and e^* satisfies $c(e^*) = 3(H - L)$ 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

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

The transfer equals $s_t = c(e^*) - (H - L)$ if agent t does not deviate and $s_t = -(H - L)$ if he does. If either nobody deviates or agent t deviates from (e_t, μ_t) but then nobody deviates

from s_t, then $a_t = H$; otherwise, $a_t = L$. 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

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

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

The principal earns $e^* - c(e^*) + H + (H-L)$ surplus in each period of the cooperative phase. Denote $s_t^P \geq 0$ and $s_t^A \geq 0$ as the principal's and agent t's transfer to each other, respectively, so that $s_t = s_t^P - s_t^A$. If agent t deviates in (e_t, μ_t), then he earns $s_t^P + L$ by paying $s_t^A = (H-L)$ and $s_t^P - s_t^A + L$ from deviating, so he has no profitable deviation from s_t^A. Regardless of μ_t, the principal has no profitable deviation from $s_t^P = 0$ following a deviation in (e_t, μ_t) if

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

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

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

This is because, by definition of α,

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

If $s_t = c(e^*) - (H-L) < 0$, then agent t has no profitable deviation from it because $c(e^*) - (H-L) + H \geq L$.

Given these transfers, agent t earns L from choosing the equilibrium (e_t, μ_t) and no more

than 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.7 Proof of Proposition 8

In the extortion game, this result follows immediately from the fact that agents are indifferent among messages and so are willing to follow their communication protocols. 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. ■

A.8 Proof of Proposition 9

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. ■

B Online Appendix: Long-run Agents

B.1 A Result with long-run Agents

B.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.[12] We also restrict attention to equilibria in pure strategies to simplify agents' beliefs on the equilibrium path.

[12] 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.

Proposition 10 *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*

$$(1-\delta)c(e_i^*) \leqslant 2\delta\rho_i s_i^* + \frac{2\rho_i\delta}{1-(1-\rho_i)\delta}\sum_{j\neq i}\rho_j s_j^*. \tag{18}$$

It is instructive to compare the right-hand side of (18) to the condition $c(e^*) \leq 3(H-L)$ from Proposition 7. 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 7. In Proposition 7, 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 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 (18).

The second term on the right-hand side of (18) represents a new force for cooperation that is not present in Proposition 7. Since each agent i chooses a new communication protocol 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 10 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 7: 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 10 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.

B.1.2 Proof of Proposition 10

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

$$s^* \leqslant \frac{\delta}{1-\delta}\left(\overline{\Pi}^* - \underline{\Pi}^*\right). \tag{19}$$

Second, the active agent i must be willing to choose effort e and the equilibrium communication protocol 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

$$s^* - c(e) + \frac{\delta}{1-\delta}\overline{U}_i^* \geqslant \max\left\{0, \frac{\delta}{1-\delta}\left(\overline{\Pi}^{HU} - \underline{\Pi}^{HU}\right)\right\}, \tag{20}$$

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

Combining (19) and (20) yields the following necessary condition for effort e to be part of equilibrium:

$$c(e) \leqslant \frac{\delta}{1-\delta}\left(\overline{U}_i^* + \overline{\Pi}^* - \underline{\Pi}^*\right) - \max\left\{0, \frac{\delta}{1-\delta}\left(\overline{\Pi}^{HU} - \underline{\Pi}^{HU}\right)\right\} \tag{21}$$

Our next goal is to connect (19) and (20) 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

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

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 $\hat{h}^{t+1}$,

$$\mathbb{E}\left[\Pi_{t+1} | \hat{h}^{t+1}\right] = \sum_{t'=t+1}^{\infty} (1-\delta)\delta^{t'-t-1} \left(\mathbb{E}\left[\pi_{i,t'} | \hat{h}^{t+1}\right] + \sum_{\tilde{t}=t+1}^{t'} \mathbb{E}\left[\pi_{-i,t'} | \hat{h}^{t+1}, \hat{\mathcal{B}}^{\tilde{t}}\right] \Pr\left\{\hat{\mathcal{B}}^{\tilde{t}}\right\} \right). \tag{22}$$

The right-hand side of (22) is absolutely convergent, so we can rearrange the order of summation to yield

$$\mathbb{E}\left[\Pi_{t+1} | \hat{h}^{t+1}\right] = \begin{array}{c} \sum_{t'=t+1}^{\infty} (1-\delta)\delta^{t'-t-1} \mathbb{E}\left[\pi_{i,t'} | \hat{h}^{t+1}\right] + \\ \sum_{\tilde{t}=t+1}^{\infty} \left(\sum_{t'=t+1}^{\tilde{t}-1} (1-\delta)\delta^{t'-t-1} \mathbb{E}\left[\pi_{-i,t'} | \hat{\mathcal{B}}^{\tilde{t}}\right] + \delta^{\tilde{t}-t-1} \mathbb{E}\left[\Pi_{-i,\tilde{t}} | \hat{\mathcal{B}}^{\tilde{t}}\right] \right) \Pr\left\{\hat{\mathcal{B}}^{\tilde{t}}\right\}. \end{array} \tag{23}$$

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 (23) to bound the difference

$$\begin{gathered}\mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] - \mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] \leqslant \\ \textstyle\sum_{t'=t+1}^{\infty} \delta^{t'-t-1}(1-\delta)\left(\mathbb{E}\left[\pi_{i,t'}|h^{t+1}\right] - \mathbb{E}\left[\pi_{i,t'}|\hat{h}^{t+1}\right]\right) + \\ \textstyle\sum_{\tilde{t}=t+1}^{\infty} \delta^{\tilde{t}-t-1}\left(\mathbb{E}\left[\Pi_{-i,\tilde{t}}|\mathcal{B}^{\tilde{t}}\right] - \mathbb{E}\left[\Pi_{-i,\tilde{t}}|\hat{\mathcal{B}}^{\tilde{t}}\right]\right)\Pr\left\{\mathcal{B}^{\tilde{t}}\right\}\end{gathered} \tag{24}$$

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 (24) is therefore

$$\begin{gathered}\mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] - \mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] \leqslant \\ \textstyle\sum_{t'=t+1}^{\infty} \delta^{t'-t-1}\left((1-\delta)\mathbb{E}\left[\pi_{i,t'}|h^{t+1}\right] + \mathbb{E}\left[\Pi_{-i,t'}|\mathcal{B}^{t'}\right]\Pr\left\{\mathcal{B}^{t'}\right\}\right)\end{gathered} \tag{25}$$

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 (25) requires

$$\mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] - \mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] \leqslant \sum_{t'=t+1} \delta^{t'-t-1}\left((1-\delta)\rho_i s_i^* + \sum_{j\neq i}\rho_j s_j^* \Pr\left\{\mathcal{B}^{t'}\right\}\right). \tag{26}$$

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 (26) again holds.

Since $s_j^* \geqslant 0$, the right-hand side of (26) 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 (26) requires

$$\begin{array}{rcl} \mathbb{E}\left[\Pi_{t+1}|h^{t+1}\right] - \mathbb{E}\left[\Pi_{t+1}|\hat{h}^{t+1}\right] & \leqslant & \sum_{t'=t+1}\delta^{t'-t-1}\left((1-\delta)\rho_i s_i^* + (1-\rho_i)^{t'-t-1}\rho_i\sum_{j=1}^N \rho_j s_j^*\right) \\ & = & \rho_i s_i^* + \frac{\rho_i}{1-(1-\rho_i)\delta}\sum_{j\neq i}\rho_j s_j^*. \end{array} \tag{27}$$

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 (27), 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 (21) 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. ■

C Online Appendix: Communication by the Principal

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 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 the extortion game (i.e., Proposition 2). The principal's message does not mitigate extortion at all, so our impossibility result still holds.

Proposition 11 *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.[13] The same argument as in the previous case applies, so every equilibrium involves zero effort in each period.

C.2 The Principal Can Commit to a Communication Protocol

In this appendix, we modify the extortion game by allowing the principal to choose a communication protocol 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

[13] 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.

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 communication protocol $\nu_t : \mathbb{R} \to M$ 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.

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 communication protocols. 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

$$c(e^*) = \frac{\delta}{1-\delta}\gamma\left(e^* - c(e^*)\right) \tag{28}$$

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 (28) 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 communication protocol eliminates extortion. Essentially, the principal's and each agent's communication protocols 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 5, 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 communication protocol 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 communication protocol as an equilibrium refinement; no analogue to Proposition 8 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.

D Online Appendix: Variants of the Extortion Game

D.1 Up-Front Transfers

The **extortion game with up-front transfers** is identical to the extortion game except that at the start of each period t, the principal and agent t exchange nonnegative transfers. Denote the resulting net wage to agent t by $w_t \in \mathbb{R}$, so that the principal's and agent t's payoffs are $e_t - s_t - w_t$ and $s_t + w_t - c(e_t)$, respectively. Note that agent t chooses a communication mechanism μ_t only *after* these transfers are paid.

Proposition 12 *Any equilibrium of the extortion game with up-front transfers entails* $e_t = s_t = 0$.

Proof of Proposition 12

Borrowing notation from the proof of Proposition 2, $s_t \leq \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi})$ on the equilibrium path, and any $s_t < \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi})$ can be made a unique best response after agent t deviates. Therefore, $c(e_t) = 0$ in any t of any equilibrium. ■

D.2 *Ex Ante* Extortion

If the principal and each agent can exchange up-front transfers, as they do in appendix D.1, it is natural to consider equilibria if agents can commit to their communication protocols as a function of those transfers. In particular, an agent might demand an up-front transfer in exchange for refraining from later extortionary threats. Of course, once the principal pays an up-front transfer to an agent, that agent has every incentive to renege on her earlier promise not to engage in extortion. In this section, we prove that even if agents can commit to not engage in future extortion in exchange for up-front payments, the unique equilibrium outcome still entails zero effort in each period.

To make this point, consider the follow game. The **game with *ex ante* extortion** is identical to the extortion game with up-front transfers, except that at the start of each

period t, agent t chooses a **communication meta-protocol** $\mu_t^0 : \mathbb{R}^2 \to \mathcal{M}$, where

$$\mathcal{M} \equiv \{\mu : \mathbb{R} \to M\}$$

is the set of communication protocols. This meta-protocol is observed by the principal but not by other agents. The principal and agent t then exchange up-front transfers, with net transfer w_t, and agent t chooses $e_t \geq 0$. Agent t's communication protocol equals the one specified by the meta-protocol, given (w_t, e_t):

$$\mu_t^0(w_t, e_t)(\cdot).$$

The rest of the period proceeds with this communication protocol.

While this alternative game might seem cumbersome, it is designed to capture a very simple intuition. In the extortion game, extortion involves shirking and so is inefficient. In principle, an agent could more efficiently extort the principal by demanding an up-front transfer in exchange for refraining from further extortion. Clearly, it might be difficult for an agent to commit to not make future extortionary threats. Even if he can overcome this commitment problem, however, *ex ante* extortion does not facilitate cooperation. In particular, each agent can use *ex ante* extortion to demand the entire proceeds from his effort. But then the principal earns nothing in any period, which means that she is unwilling to compensate any agent for his effort. The resulting unique equilibrium outcome entails no effort.

Proposition 13 *Every equilibrium of the game with* ex ante *extortion entails* $e_t = 0$ *in each* $t \geq 0$.

Proof of Proposition 13

Define $\Pi^* \geq 0$ as the principal's maximum equilibrium payoff, and consider an equilibrium that attains Π^*. If $e_0 = 0$ with probability 1 in this equilibrium, then

$$\Pi^* \leq \delta \Pi^*$$

and so $\Pi^* = 0$.

Suppose that $e_0 > 0$ with positive probability in this equilibrium. Define $\bar{\Pi}$ and $\underline{\Pi}$ as the largest and smallest equilibrium continuation payoffs induced by m_0 in this equilibrium, with corresponding messages C and D, respectively. Let e^* equal the effort that maximizes total period-0 surplus among all on-path efforts in period 0. Then $e^* > 0$, and moreover, it must be that

$$\frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}) \geq c(e^*) > 0.$$

Fix some $\hat{w} \geq 0$. For any $\epsilon, \xi > 0$, consider the following choice of μ_0^0 by agent 0:

$$\mu_0^0(w_0, e_0)(\cdot) = \begin{cases} \mu^C(\cdot) & \text{if } e_t = e^* - \epsilon, w_t = \hat{w} \\ \mu^C(\cdot) & \text{if } e_t = 0, w_t \neq \hat{w} \\ \mu^D(\cdot) & \text{otherwise} \end{cases},$$

where

$$\mu^C(s_0) = \begin{cases} C & s_0 = \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}) - \xi \\ D & \text{otherwise} \end{cases}$$

and

$$\mu^D(s_0) = D.$$

Give this choice, suppose that $w_0 \neq \hat{w}$. If $e_0 = 0$, then the principal's unique best

response (to μ^C) is

$$s_0 = \frac{\delta}{1-\delta}\left(\bar{\Pi} - \underline{\Pi}\right) - \xi.$$

If $e_0 > 0$, then the principal's unique best response (to μ^D) is $s_0 = 0$. Therefore, agent t's uniquely optimal effort is $e_0 = 0$, in which case the principal's payoff is at most

$$(1-\delta)\xi + \delta\underline{\Pi}.$$

Suppose instead that $w_0 = \hat{w}$. If $e_0 = e^* - \epsilon$, then the principal's unique best response (to μ^C) is again

$$s_0 = \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}) - \xi.$$

If $e_0 \neq e^* - \epsilon$, then the principal's unique best response (to μ^D) is $s_0 = 0$. For any $\epsilon > 0$, there exists a sufficiently small $\xi > 0$ such that

$$-c(e^* - \epsilon) + \frac{\delta}{1-\delta}\left(\bar{\Pi} - \underline{\Pi}\right) - \xi > 0.$$

Therefore, $e_0 = e^* - \epsilon$ is agent 0's uniquely optimal effort, in which case the principal's payoff is

$$(1-\delta)(\xi + e^* - \epsilon - \hat{w}) + \delta\underline{\Pi}.$$

We have uniquely pinned down the principal's payoff as a function of $\hat{w}$. The principal's finds it strictly optimal to pay $w_0 = \hat{w}$ so long as

$$(1-\delta)(\xi + e^* - \epsilon - \hat{w}) + \delta\underline{\Pi} > (1-\delta)\xi + \delta\underline{\Pi}$$

or

$$\hat{w} < e^* - \epsilon.$$

Agent 0 can therefore use this strategy to guarantee a payoff arbitrarily close to

$$e^* - \epsilon - c(e^* - \epsilon) + \frac{\delta}{1-\delta}(\bar{\Pi} - \underline{\Pi}) - \xi. \tag{29}$$

In equilibrium, agent 0's utility and the principal's payoff cannot exceed

$$(1-\delta)(e^* - c(e^*)) + \delta\bar{\Pi}.$$

Subtracting (29) from this surplus and taking $\epsilon, \xi \to 0$, we conclude that the principal's payoff cannot exceed $\delta\underline{\Pi}$. But then

$$\Pi^* \leq \delta\underline{\Pi} \leq \delta\Pi^*,$$

so again $\Pi^* = 0$.

We have established that the principal's *maximum* equilibrium payoff equals 0, which is also her min-max payoff. Therefore, $\bar{\Pi} = \underline{\Pi} = 0$, which implies that $e_t = 0$ in each $t \geq 0$ of any equilibrium. ■

While every equilibrium entails zero effort in the game with *ex ante* extortion, the intuition for this result differs from that of Proposition 2. Here, each agent can use the meta-communication protocol to extract the entire surplus created by his interaction with the principal. The principal therefore has no reason to actually pay an agent, since her continuation payoff equals zero regardless of her actions today. Proposition 13 is a particularly extreme consequence of the negative intertemporal externality from Proposition 4. In this case, each agent's rent-seeking behavior is so severe that it totally undermines cooperation, resulting in zero effort in equilibrium.

This model assumes that agent t's meta-protocol conditions on both the up-front transfer and his effort. We make this assumption to draw a close connection to the extortion model, since agents in that model implicitly condition their communication protocols on their efforts.

Note that assuming μ_t^0 can condition on e_t does not resolve the commitment problem at the heart of the model, since agent t must still find it sequentially optimal to exert effort given his beliefs about s_t. We could instead assume that μ_t^0 can condition on w_t but *not* on e_t, in which case we can construct equilibria with strictly positive effort. These equilibria take advantage of the fact that once an agent deviates in μ_t^0, there might exist a continuation equilibrium in which he expects $s_t = 0$ and so exerts no effort. Analogous to Proposition 7, the possibility of inefficient continuation play *within* period t limits agent t's ability to engage in *ex ante* extortion. The principal can therefore earn a strictly positive equilibrium payoff if meta-protocols are not contingent on effort.