

# The Use and Abuse of Coordinated Punishments

Daniel Barron and Yingni Guo\*

October 22, 2018

## Abstract

Communication facilitates cooperation by allowing deviators to be widely punished. This paper explores how players might misuse communication to threaten one another, and it studies conditions under which cooperation is possible despite these threats. A principal plays trust games with a sequence of short-lived agents who communicate. An agent who shirks can still demand payment by threatening to report that the principal deviated. We show how these threats can destroy cooperation. Cooperation is partially restored if agents sometimes fail to follow through on threats, or the principal can credibly punish shirking agents, or agents have hard evidence about their pay.

---

\*Barron: Northwestern University Kellogg School of Management, Evanston IL 60208; email: d-barron@kellogg.northwestern.edu. Guo: Northwestern University Economics Department, Evanston IL 60208; email: yingni.guo@northwestern.edu. The authors would like to thank Nemanja Antic, Joyee Deb, Matthias Fahn, George Georgiadis, Marina Halac, Peter Klibanov, Ilan Kremer, Nicolas Lambert, Stephan Lauermann, Jin Li, Elliot Lipnowski, Bentley MacLeod, David Miller, Joshua Mollner, Michael Powell, Luis Rayo, Mark Satterthwaite, Takuo Sugaya, and Jeroen Swinkels.

# 1 Introduction

The threat of sanctions motivates individuals and firms to work hard and reward the hard work of their partners (Malcomson [2013], Gibbons and Henderson [2013]). Word-of-mouth can be an essential tool for carrying out these sanctions, as news of misbehavior spreads far beyond those who directly observed that misbehavior. Communication therefore lies at the heart of many cooperative endeavors, from managers who reward the efforts of their employees because they otherwise fear a widespread strike (Freeman and Lazear [1995], Levin [2002]), to communities that ostracize members who do not contribute to public goods (Ostrom [1990], Ali and Miller [2016]), to business associations that limit opportunistic behavior by spreading word of past misdeeds (Greif et al. [1994], Robinson and Stuart [2006], Bernstein [2015]), to firms that provide exemplary service to avoid negative reviews and lost sales (Hörner and Lambert [2017]).

Communication facilitates cooperation in these settings by allowing parties who do not observe a transgression to nevertheless punish the transgressor. Once armed with messages that trigger punishment, however, parties might misuse those messages to extract concessions from one another. For example, even a worker who shirks might be able to demand undeserved rewards by threatening to otherwise complain to his union.<sup>1</sup> Similarly, influential members of a community can threaten to ostracize those who refuse to comply with the existing social order,<sup>2</sup> members of business associations can threaten to falsely report misbehavior unless their partners agree to excessively generous terms,<sup>3</sup> and customers can threaten to leave negative reviews unless a business inefficiently favors them.<sup>4</sup>

---

<sup>1</sup>We are far from the first to suggest that unions, and the coordination they facilitate, can be used to pursue both welfare-improving and welfare-destroying goals. See, for instance, Freeman and Medoff [1979] and Freeman [1980].

<sup>2</sup>In extreme cases, this ostracism can result in outright persecution. For example, some scholars have argued that the witch-hunt craze in 14th-17th century Europe, which was responsible for as many as half a million deaths, was a reaction by old authority structures to new social norms (Ben-Yehuda [1980]).

<sup>3</sup>Businesses sometimes “blacklist” employees, which prevents them from seeking employment from other companies in the industry. NPR’s Planet Money podcast has reported allegations that Wells Fargo misused a tool designed to share information about unethical workers to blacklist former employees who protested fraudulent sales practices (Arnold and Smith [2016]).

<sup>4</sup>For example, Klein et al. [2006] and Klein et al. [2017] argue that prior to a change to Ebay’s rating

In this paper, we study how rent-seeking agents can misuse communication systems that are designed to facilitate cooperation, and we explore how cooperation can be sustained in the face of these threats. We consider a simple model in which (i) cooperation depends on communication, but (ii) players who deviate can still threaten to send messages that trigger punishment unless their partners reward them. We begin with a stark negative result: these threats totally undermine cooperation, as any attempt at punishing a deviation simply becomes an opportunity for parties to extract resources without exerting effort. We then build on this negative result to identify settings in which cooperation is possible despite these threats. Collectively, our results both suggest that coordinated punishments are susceptible to abuse and identify features of the environment that might ameliorate that abuse and restore cooperation.

To make these points, we study a long-lived principal who interacts with a sequence of short-lived agents. Each agent exerts costly effort to benefit the principal, who can then pay that agent. Agents observe only their own interaction with the principal but can costlessly communicate. Our key modeling ingredient is that an agent who deviates can still threaten to tailor his report to the principal's payment. To make this threat credible, we assume that each agent chooses a **communication protocol** at the start of his interaction with the principal. This protocol, which is observed by the principal, associates a message to each possible payment. The agent is then committed to communicate according to his protocol at the end of that period.

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 is otherwise punished. Communication is therefore essential for inducing effort, since agents cannot directly punish the principal for renegeing. Once armed with messages that can trigger punishment, however, an agent can shirk and then demand payment using the same threats as an agent who exerted effort. Since

---

system, reviews were largely uninformative because sellers could deter buyers from writing negative reviews by threatening to retaliate with reciprocal negative reviews. Similarly, the Canadian Broadcasting Company covered a recent lawsuit against a disgruntled customer who used defamatory online reviews to drive a photography studio out of business (Proctor [2018]).

these threats are enough to induce the principal to compensate an agent for her efforts, they are also enough to force 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. The resulting unique equilibrium outcome entails no payments and no effort.

This (intentionally stark) negative result forms the foundation for the rest of the paper, which studies three changes in the environment that remedy this problem and lead to at least some effort in equilibrium. Our first remedy considers equilibrium outcomes if agents are only probabilistically committed to their communication protocols. Agents can still make threats in this setting, but those threats are not fully credible and so some effort can be sustained in equilibrium. The higher the probability that an agent is committed to his communication protocol, the more tempting is his deviation to shirk and threaten to report defection unless he is paid. An agent is deterred from deviating in this way only if he is paid a rent. But paying rent to an agent creates a negative intertemporal externality on the principal's relationships with other agents, since it means that the principal earns lower continuation surplus and so is less willing to pay large rewards for effort. This intuition extends to the entire equilibrium payoff set, which shrinks as the probability of commitment increases and includes only the one-shot equilibrium payoff if agents are always committed.

Our second remedy introduces hard evidence about the principal's payment. Evidence means that an agent cannot lie about the principal's payments, which limits the kinds of threats he can make and so leads to more cooperation. This mechanism is a little subtle, however, because non-payment generates the same evidence regardless of whether an agent works hard or shirks. In equilibrium, we show that evidence is used to make the principal indifferent between payments, so that she is willing to both pay a hard-working agent and not pay an agent who shirks and threatens her. For the principal to be indifferent, the *absence* of evidence must be treated as harshly as evidence of malfeasance. Consequently, if evidence is only probabilistically available, then the principal is periodically punished on the equilibrium path. We also show that this indifference condition can be impossible to satisfy

if evidence is noisy, in which case no cooperation is possible in equilibrium.

Our final remedy introduces a bilateral relationship between the principal and each agent. We model this relationship in an abstract, “reduced form” way by allowing each agent to play a coordination game with the principal *after* he has communicated with the other agents. This coordination game can be used to encourage cooperation by directly punishing the principal or an agent for deviating. In addition, we show that bilateral relationships enable coordinated punishments, since the coordination game can also be used to punish agents for trying to extract rent and the principal for giving in to those threats. However, the strength of the bilateral relationship tightly constrains how severe coordinated punishment can be while remaining effective. In Appendix C, we enrich this intuition by considering a setting with long-lived agents who interact repeatedly with the principal and show how these bilateral interactions enable at least some coordinated punishments.

The essential assumption in our model is that agents can commit to their messages, which means that an agent who has shirked can still credibly threaten to spread word that the principal deviated. We show that this commitment assumption selects equilibria of the game without commitment, in the sense that our equilibria are outcome-equivalent to equilibria if the agents cannot commit. In other words, commitment is powerful in our setting because it allows agents to specify their messages in advance, not because it results in messages that agents would otherwise be unwilling to send. We also argue that in the baseline game and the game with evidence, agents act as if they are committed so long as they have a mild intrinsic preference for following through on their communication protocols.

The broader lesson of our analysis is that the very features that make communication valuable also make it susceptible to abuse by rent-seeking agents. Communication is valuable if parties (i) do not directly observe one another’s relationships, and (ii) cannot, on their own, severely punish deviations in their own relationships. These features mean that parties can threaten one another with little fear of being caught and at low personal cost. These threats have the potential to severely limit equilibrium cooperation, though features of the

environment can mitigate the resulting inefficiencies.

## Related Literature

Our paper complements the literature on how communication improves cooperation in settings without formal commitment. The classic study by Greif et al. [1994] shows how institutions can facilitate communication and deter deviations. Ali et al. [2016], which analyzes renegotiation-proof equilibria in settings where players must communicate about deviations, is perhaps the most closely related paper in this literature. That paper shows that renegotiation does not inhibit cooperation if communication is costless. We consider hold-up rather than renegotiation, and we show that, unlike renegotiation, the threat of hold-up can severely limit equilibrium cooperation. Much of this literature focuses on *networks* of players with the goal of identifying network structures or equilibrium strategies that are particularly conducive to cooperation (Lippert and Spagnolo [2011]; Wolitzky [2013]; Ali and Miller [2016]; Ali and Liu [2018]). Rather than focusing on network structure, we instead focus on how communication can be abused by rent-seeking agents. We introduce a new ingredient, the communication protocol, to ensure that agents' hold-up threats are credible. More distantly related papers 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 and communication.

Commitment allows our agents to extract rent by threatening the principal. As we will show, these threats can be interpreted as an equilibrium refinement, which links our paper to the relatively small literature on bargaining in repeated games. For instance, several papers impose the constraint that players Nash bargain over either surplus in each period or continuation play (Baker et al. [2002], Halac [2012], Halac [2015], Goldlucke and Kranz [2017]). The most closely related papers in this literature are Miller and Watson [2013] and Miller et al. [2018], which define and analyze *contractual equilibria* in relational contracts. Contractual equilibria allow players to propose continuation equilibria to one

another, which both incorporates a notion of renegotiation-proofness and limits how surplus can be split among parties. Our model of hold-up leads to a different and complementary set of constraints. In particular, our equilibria are not renegotiation-proof and so can entail harsh punishments. However, harsher punishments have a downside, since they give the agents more substantial opportunities to hold up the principal. Our analysis also focuses on communication among agents, while these papers mostly emphasize frictions that can arise in either bilateral relationships or settings with public monitoring.

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 the costs and benefits of communication among multiple agents. Recent papers have explored relational contracts with additional frictions in the environment, including 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 different friction, bilateral monitoring within each principal-agent pair, which implies that communication is essential for cooperation. Other papers that study settings 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 a relatively ineffective tool for sustaining cooperation.

Levin [2002], which studies relational contracts between a principal and a group of agents, is the seminal paper within this literature that touches on both coordinated punishments and bargaining. Proposition 5 of that paper touches on the idea of rent-seeking leading to a negative intertemporal externality, which is also a feature of our setting. In other respects, however, both our central mechanism and its implications differ substantially from Levin [2002]: we emphasize how communication can be abused in settings with private (rather than public) monitoring, and we identify features of the environment that mitigate this vulnerability.

Our assumption that agents commit to messages is similar to the persuasion literature, as in Rayo and Segal [2010] and Kamenica and Gentzkow [2011]. Particularly related are Min [2017], Lipnowski et al. [2018], and Guo and Shmaya [2018], which similarly assume that players are not perfectly committed to their messages. However, commitment serves a very different role in our setting, since it allows agents to make off-path threats rather than reveal information about payoff-relevant information. Consequently, agents’ equilibrium payoffs are non-monotonic in the probability of commitment in our setting.<sup>5</sup>

## 2 Model

A long-lived principal (“she”) interacts with a sequence of short-lived agents (each “he”). In each period  $t \in \{0, 1, 2, \dots\}$ , the principal and agent  $t$  play a favor-trading game: agent  $t$  exerts effort and exchanges payments with the principal. Agents do not observe one another’s interactions with the principal but can costlessly communicate. The key assumption is that before transfers are paid, each agent chooses a *communication protocol*, which is a mapping from the net transfer to this agent’s message. An agent is committed to communicate according to this protocol with probability  $\lambda \in [0, 1]$  and is otherwise free to choose any message.

We assume that players have access to a public randomization device (notation for which is suppressed) in every step of the stage game. The stage game in each period  $t$  is:

1. Agent  $t$  chooses his effort  $e_t \in \mathbb{R}_+$  and a communication protocol  $\mu_t : \mathbb{R} \rightarrow M$ , where  $M$  is a large, finite message space. Both  $e_t$  and  $\mu_t$  are observed by the principal but not by any other agent.
2. The principal and agent  $t$  simultaneously pay non-negative transfers to one another, which are observed by those two players but not by any other agent. Let  $s_t \in \mathbb{R}$  be

---

<sup>5</sup>This non-monotonicity arises because neither effort nor payments are contractible and so is a particular manifestation of the Theory of the Second Best (Lipsey and Lancaster [1956-1957], Bernheim and Whinston [1998]).

the net transfer to agent  $t$ .<sup>6</sup>

3. Agent  $t$  sends a message  $m_t$  that is publicly observed. With probability  $\lambda \in [0, 1]$ , agent  $t$  is committed to  $\mu_t$  so  $m_t = \mu_t(s_t)$ ; otherwise, agent  $t$  chooses any  $m_t \in M$ .

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 differentiable, strictly increasing, strictly convex, and satisfies  $c(0) = c'(0) = 0$ . We assume that there exists a unique effort level, denoted  $e^{FB}$ , that solves  $c'(e^{FB}) = 1$  and so maximizes total surplus. The principal has discount factor  $\delta \in [0, 1)$ , with corresponding normalized discounted payoffs  $\Pi_t = (1 - \delta) \sum_{t'=t}^{\infty} \delta^{t'-t} (e_{t'} - s_{t'})$ . We sometimes consider the normalized discounted sum of agents' utilities,  $U_t = (1 - \delta) \sum_{t'=t}^{\infty} \delta^{t'-t} (s_{t'} - c(e_{t'}))$ , which we call the agents' *joint utility*.

Our solution concept is plain Perfect Bayesian Equilibrium (hereafter PBE) as defined in Watson [2016]. Many of our results focus on *principal-optimal equilibria*, which maximize the principal's *ex ante* expected payoff  $\mathbb{E}[\Pi_0]$  among PBE.

Our key modeling ingredient is that agents can commit to a communication protocol, which means that they can credibly promise to send messages that reward or punish the principal. In particular, agents can abuse communication by shirking and then threatening to punish the principal unless she pays them. Section 5 explores this assumption in more detail.

The assumption that agents are short-lived lets us cleanly highlight how communication can lead to hold-up. Even with long-lived agents, however, communication remains both valuable and susceptible to hold-up. Section 4.2 shows how future bilateral relationships, modeled as a coordination game played between the principal and each agent at the end of the period, can mitigate hold-up. Appendix C considers a repeated game with long-lived agents and shows how the threat of hold-up undermines cooperation in that setting.

Several other assumptions warrant further comment. First, we do not include a round

---

<sup>6</sup>With the exception of Section 4.2, the agents do not pay the principal and so requiring  $s \geq 0$  would not change the results.

of transfers at the start of each period. Appendix D.1 shows that allowing such transfers would not change the principal-optimal equilibrium in our baseline model, since requiring an agent to make an up-front payment simply gives them a stronger incentive to deviate and hold up the principal. Second, our model allows the agents but not the principal to send public messages. Appendix D.2 shows that even if the principal can send messages, our results extend if either  $\lambda = 1$  or the principal chooses her message after observing an agent's message. Finally, we do not allow agents to interact directly. Appendix D.3 allows each agent to interact with his immediate predecessor. These inter-agent interactions do not improve cooperation, provided that they precede the principal's interaction with the agent. In that case, each agent cannot infer, and so cannot reward or punish, his predecessor's effort or communication protocol.

### 3 Multilateral Punishments Enable Hold-Up

This section shows how hold-up undermines cooperation. Each agent is tempted to shirk and then use the communication protocol to induce the principal to nevertheless pay him. If  $\lambda = 1$ , this temptation leads to the complete unravelling of cooperation and agents shirk in every equilibrium. If  $\lambda < 1$ , then the principal deters the agents from deviating by paying them rent, which limits the credibility of her promises to other agents and decreases effort. We build intuition using the full- and no-commitment cases before proving a general characterization.

Cooperation requires informative communication among the agents, since without communication an agent would have no way to punish the principal for reneging on a payment and so would not exert effort. Our first result shows that communication can indeed sustain cooperation if agents cannot commit to their communication protocols.

**Proposition 1.** *Suppose that  $\lambda = 0$ . Let  $e^*$  equal the minimum of  $e^{FB}$  and the positive root of  $c(e) = \delta e$ . In any principal-optimal equilibrium,  $e_t = e^*$  and  $s_t = c(e^*)$  in every  $t \geq 0$ .*

**Proof:** Let  $\bar{\Pi}$  be the principal's maximum equilibrium payoff, and let  $e^*$  be the maximum equilibrium effort in an principal-optimal equilibrium. The principal pays  $s_t$  only if  $(1-\delta)s_t \leq \delta\bar{\Pi}$ , while agent  $t$  chooses  $e_t = e^*$  only if  $s_t - c(e^*) \geq 0$ . Effort  $e^*$  must therefore satisfy  $(1-\delta)c(e^*) \leq \delta\bar{\Pi}$ . If this necessary condition is sufficient, then total surplus would be maximized by  $\bar{\Pi} = e^{FB} - c(e^{FB})$  whenever  $c(e^{FB}) \leq \delta e^{FB}$ . Otherwise,  $e^*$  would be bounded above by  $(1-\delta)c(e^*) = \delta\bar{\Pi}$ , which implies  $c(e) = \delta e$  because  $\bar{\Pi} = e^* - c(e^*)$ .

Next, we prove that this necessary condition is sufficient in a principal-optimal equilibrium. Consider the following strategy: in each period  $t \geq 0$  on path,  $e_t = e^*$ ,  $s_t = c(e^*)$ , and  $m_t = m^*$ . If agent  $t$  deviates in  $e_t$ , then  $s_t = 0$  and  $m_t = m^*$ . If the principal deviates, then  $m_t \neq m^*$ . Once a message  $m_t \neq m^*$  is observed, all future periods entail repetition of the stage-game equilibrium  $e_t = s_t = 0$ . Agents are indifferent among messages and so are willing to send the specified messages. The principal is willing to pay  $s_t = c(e^*)$  because  $(1-\delta)c(e^*) \leq \delta(e^* - c(e^*))$ . Agent  $t$  is indifferent between choosing  $e_t = e^*$  and  $e_t = 0$ . We conclude that this strategy is a principal-optimal equilibrium, since it maximizes total surplus and gives that entire surplus to the principal. ■

An agent exerts effort in equilibrium only if he expects a high payment for working and a low payment for shirking. Proposition 1 shows that communication can induce the principal to pay agents who work hard, which is enough to sustain effort in a setting without commitment. Equilibrium effort is bounded from above by the dynamic enforcement constraint  $\delta e \leq c(e)$ . Note that many other equilibria also exist; for example, if  $\delta > c(e^{FB})/e^{FB}$ , then equilibria exist that induce first-best effort but give the agents strictly positive utility.

If  $\lambda > 0$  then the coordinated punishments used in the proof of Proposition 1 leads to a new problem: by choosing a communication protocol that punishes the principal for non-payment, an agent can guarantee a positive transfer even if he shirks. We refer to this deviation, in which an agent shirks and then threatens to punish the principal unless she pays him, as hold-up. Our next result shows that if  $\lambda = 1$ , then the possibility of hold-up means that an agent can earn essentially the same payment regardless of his effort. Consequently,

agents never exert effort in equilibrium.

**Proposition 2.** *If  $\lambda = 1$ , then every equilibrium entails  $e_t = s_t = 0$  in each  $t \geq 0$ .*

**Proof:** Suppose that  $e_t > 0$  in equilibrium. Define  $\bar{\Pi}$  and  $\underline{\Pi}$  as the principal's maximum and minimum continuation payoffs in  $t + 1$  onwards, and let  $\bar{m}_t$  and  $\underline{m}_t$  be the corresponding messages that lead to these payoffs. Since  $e_t > 0$ ,  $s_t \geq c(e_t) > 0$  and hence  $\bar{\Pi} > \underline{\Pi}$ .

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

$$\mu_t(s) = \begin{cases} \bar{m}_t & s \geq s_t - \varepsilon \\ \underline{m}_t & \text{otherwise.} \end{cases}$$

The principal's unique best response to this deviation is to pay  $s_t - \varepsilon$ , since

$$-(s_t - \varepsilon) + \frac{\delta}{1 - \delta} \bar{\Pi} > \frac{\delta}{1 - \delta} \underline{\Pi}$$

is implied by the fact that the principal is willing to pay  $s_t$  on the equilibrium path. For any  $\varepsilon \in [0, c(e_t))$ ,  $s_t - \varepsilon > s_t - c(e_t)$  and so this deviation is profitable for agent  $t$ . We conclude that no equilibrium can entail  $e_t > 0$ . But if  $e_t = 0$  for all  $t \geq 0$ , then  $s_t = 0$  as well. ■

If  $s_t > 0$  on the equilibrium path, then agent  $t$  can choose zero effort 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 is strictly willing to pay a smaller amount. The agent can therefore shirk and still guarantee nearly the same transfer as if he exerted effort. Commitment essentially severs the link between payment and effort and hence eliminates the agents' incentives to work.

Propositions 1 and 2 illustrate how messages can be used to both deter the principal from renegeing on a payment and to hold her up. In equilibrium, the principal must have the incentive to both pay an agent who has worked hard and refrain from paying an agent who has shirked. As in Proposition 1, equilibrium communication enables punishments that

induce the principal to pay a working agent. As in Proposition 2, however, an agent might be able to use those same messages to force the principal to pay him even after he exerts no effort. This latter effect increases an agent's temptation to shirk such that no effort can be sustained in equilibrium if  $\lambda = 1$ .

Our final result in this section shows how principal-optimal equilibria balance these two contrasting roles for  $\lambda \in (0, 1)$ . The former role dominates for  $\lambda \leq 1/2$ , in which case equilibrium effort is the same as in the case without commitment. For higher  $\lambda$ , the latter role becomes increasingly important as  $\lambda$  increases, resulting in agents who can make more credible threats and so exert less effort relative to Proposition 1.

**Proposition 3.** *Let  $e^*(\lambda)$  equal the minimum of the effort level that solves  $c'(e) = 2(1 - \lambda)$  and the positive root of  $c(e) = 2\delta(1 - \lambda)e$ . In any principal-optimal equilibrium:*

1. *If  $\lambda \leq \frac{1}{2}$ , then  $e_t = e^*(\frac{1}{2})$  and  $s_t = c(e_t)$  in each period on the equilibrium path. The principal's payoff is constant in  $\lambda$ .*
2. *If  $\lambda > \frac{1}{2}$ ,  $e_t = e^*(\lambda)$  and  $s_t = \frac{c(e_t)}{2(1-\lambda)} > c(e_t)$  in each period on the equilibrium path. The principal's payoff is strictly decreasing in  $\lambda$ .*

**Proof:** See Appendix A.

Note that  $e^*(\frac{1}{2})$  equals the equilibrium effort from Proposition 1. To prove Proposition 3 for  $\lambda \leq \frac{1}{2}$ , it suffices to show that the principal is willing to pay nothing to an agent who deviates, because in that case the agent's incentives are identical to the case with  $\lambda = 0$ . In period  $t$ , let  $\bar{\Pi}$  and  $\underline{\Pi}$  be the principal's highest and lowest equilibrium continuation payoffs from period  $t + 1$  onwards. Suppose that, if agent  $t$  shirks and is not committed to his communication protocol, then his message leads to continuation payoff  $\bar{\Pi}$  if  $s_t = 0$  and continuation payoff  $\underline{\Pi}$  for any  $s_t > 0$ . Given these messages, the principal's continuation surplus can be no worse than  $\lambda\underline{\Pi} + (1 - \lambda)\bar{\Pi}$  if  $s_t = 0$  and no better than  $\lambda\bar{\Pi} + (1 - \lambda)\underline{\Pi}$  if  $s_t > 0$ . If  $\lambda \leq \frac{1}{2}$ , the principal's best response is  $s_t = 0$ , resulting in a principal-optimal

equilibrium that is outcome-equivalent to Proposition 1.<sup>7</sup>

By a similar logic, an agent can deviate and induce a strictly positive transfer whenever  $\lambda > \frac{1}{2}$ . Larger  $\lambda$  or  $\bar{\Pi} - \underline{\Pi}$  means that each agent can extract a larger share of the total surplus. Consequently, increasing  $\bar{\Pi} - \underline{\Pi}$  might increase effort but definitely increases the share of the output earned by agents. In the principal-optimal equilibrium, this difference is chosen to balance the value of an increase in effort against the fact that agents earn a larger share of the resulting surplus. The resulting equilibrium is either a corner solution that maximizes  $\bar{\Pi} - \underline{\Pi}$ , or it maximizes  $e - \frac{c(e)}{2(1-\lambda)}$ , which equals total surplus minus the agents' rents.

Increasing  $\lambda$  decreases total surplus but increases the agents' share of that surplus. Indeed, agents earn no rent if either  $\lambda = 0$ , in which case they cannot hold up the principal, or  $\lambda = 1$ , in which case no effort is possible in equilibrium. This intuition suggests that agents' rent is maximized at an interior probability of commitment, which is confirmed in the following example.

**Example 1.** *Suppose that  $c(e) = e^2$ . Then effort  $e = 1 - \lambda$  solves  $c'(e) = 2(1 - \lambda)$ , so  $\delta \geq \frac{c(e)}{2(1-\lambda)e}$  holds whenever  $\delta \geq \frac{1}{2}$ . If  $\delta \geq \frac{1}{2}$ , the agents' joint utility in the principal-optimal equilibrium equals  $(1 - \lambda)(2\lambda - 1)/2$ , which is maximized at  $\lambda = \frac{3}{4}$ .*

While Proposition 3 focuses on principal-optimal equilibria, a similar intuition extends to the entire equilibrium set. Indeed, Proposition 2 has already shown that the unique equilibrium outcome contains only  $e_t = s_t = 0$  if  $\lambda = 1$ . Our next result extends this argument to show that the set of equilibrium outcomes is decreasing in  $\lambda$ .

---

<sup>7</sup>One seemingly unusual feature of this construction is that an agent who deviates sometimes punishes the principal for paying him. This construction minimizes the amount the principal is willing to pay following a deviation and so minimizes an agent's incentive to deviate. However, we can sustain positive, albeit lower, effort in equilibria that do not share this feature, provided that the principal is punished less harshly for non-payment if the agent deviates than if he does not. For example, there exists an equilibrium with positive effort in which an agent who deviates and is not committed chooses *the same* message regardless of the transfer.

**Proposition 4.** *For any  $\delta \in [0, 1)$  and  $\lambda' > \lambda$ , for any equilibrium of the game with commitment probability  $\lambda'$ , there exists an equilibrium of the game with commitment probability  $\lambda$  which induces the same distribution over  $\{e_t, s_t\}_{t=0}^\infty$ .*

**Proof:** See Appendix A.

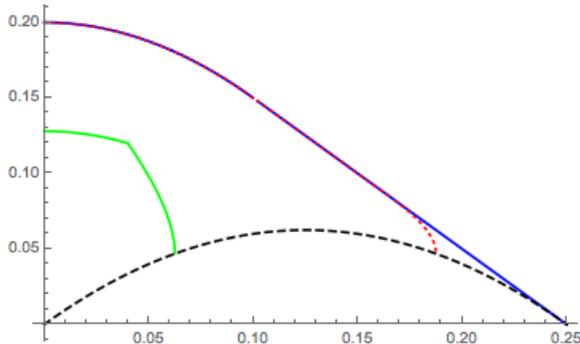


Figure 1: Payoff frontier if  $c(e) = e^2$  and  $\delta = 8/10$ . The  $x$ -axis is the principal's payoff and the  $y$ -axis is the agents' joint utility. The blue curve corresponds to  $\lambda \leq 1/2$ ; the red-dotted curve to  $\lambda = 5/8$ ; and the green curve to  $\lambda = 7/8$ . The black-dashed curve traces the principal's payoff and agents' joint utility in the principal-optimal equilibrium as  $\lambda$  increases from  $1/2$  to  $1$ , where higher  $\lambda$  correspond to points closer to the origin.

The proof of Proposition 4 follows from the fact that an agent who is not committed is willing to choose any message, including the message specified by the communication protocol. Therefore, agents are always willing to act as if they are committed to their protocols, so we can replicate the equilibrium messages from the game with high commitment probability in the game with low commitment probability. An identical mapping from transfers to messages induces identical transfers and identical efforts, proving the result.

Appendix B expands Proposition 4 by fully characterizing the equilibrium payoff frontier for any  $\lambda \in [0, 1]$ . Figure 3 illustrates how the payoff frontier changes as  $\lambda$  varies. In general, the principal-optimal equilibrium does not maximize total welfare, since making coordinated punishments more severe can increase the agents' joint utility faster than it decreases the principal's payoff. Pareto efficient equilibria are not necessarily stationary, either, though any non-stationarity occurs only in the first period of play.

Together, Proposition 4 and Appendix B illustrate a negative intertemporal externality that arises from agents' rent-seeking behavior. Increasing Agent  $t$ 's payoff means that the principal has less to lose from reneging on her promises in all periods  $t' < t$ . She is therefore less willing to reward effort, which decreases the surplus generated in these other periods. Increasing  $\lambda$  increases each agent's ability to extract rent, which exacerbates this negative externality and so decreases both effort and total surplus.

## 4 Two Remedies to the Hold-Up Problem

This section studies two imperfect solutions to the hold-up problem identified in Section 3. In Section 4.1, we introduce the possibility of hard evidence about the transfer, which limits the messages that an agent can use to threaten the principal. Section 4.2 allows each agent to play a coordination game with the principal after sending a message, which can be used to punish that agent for threatening hold-up or the principal for giving in to those threats. Both of these sections restrict attention to  $\lambda = 1$  so that the hold-up problem is particularly severe.

### 4.1 Remedy 1: Evidence

The *game with evidence* modifies the baseline game by introducing evidence about the transfer that can be concealed but cannot be distorted. After the principal pays  $s_t$ , agent  $t$  gets hard evidence of this payment with probability  $p \in [0, 1]$ . Let  $x_t \in \{\emptyset, s_t\}$  denote this evidence, where  $x_t = \emptyset$  if no evidence is available. When agent  $t$  chooses a communication protocol, he also chooses a *disclosure protocol*  $\kappa_t : \mathbb{R} \times \{\emptyset\} \rightarrow \mathbb{R} \cup \{\emptyset\}$  subject to the restriction that  $\kappa_t(x) \in \{\emptyset, x\}$  for any  $x \in \mathbb{R} \cup \{\emptyset\}$ . The disclosure protocol is observed by the principal but not by any other agents. The resulting disclosure  $k_t = \kappa_t(x_t)$  is then publicly observed by all agents.

Our main result in this section characterizes the principal-optimal equilibrium in the

game with evidence. The good news is that, unlike Proposition 2, positive effort can be attained in equilibrium. However, any such equilibrium requires the principal to be punished whenever  $k_t = \emptyset$ , which limits the gains from cooperation if  $p < 1$ .

**Proposition 5.** *Consider the game with evidence and let  $\lambda = 1$ . In any principal-optimal equilibrium, the principal's expected payoff is*

$$e^* - \frac{c(e^*)}{p},$$

where  $e^*$  is the minimum of the effort that maximizes  $e - \frac{c(e)}{p}$  and the positive root of  $\delta p e = c(e)$ .

If  $p < 1$ , then on-path play is non-stationary. There exists a principal-optimal equilibrium with the following features. Play starts in the “production” phase. In this phase,  $e_t = e^*$  and  $s_t = c(e^*)$ . If  $x_t = c(e^*)$ , play stays in the production phase. Otherwise, play moves to a “frozen” phase with strictly positive probability and otherwise stays in the production phase. The frozen phase is absorbing and features  $e_t = s_t = 0$ .

**Proof:** See Appendix A.

Suppose  $p = 1$ , in which case Proposition 5 says that the principal-optimal equilibrium in the game with evidence leads to the same effort and payoffs as in the baseline game with  $\lambda = 0$ . The logic for why it does so, however, is a little subtle. To see why, note that we would like the principal to not pay an agent who has shirked. Evidence, however, does not reveal anything about an agent's effort, and in particular it does not distinguish a principal who deviates from one who refuses to pay an agent who has deviated. However, evidence does allow the construction of an equilibrium in which the principal is *indifferent* between paying an agent and suffering punishment.

Suppose the principal's continuation payoff equals  $\Pi_{t+1}$  if the agent reveals that  $s_t = c(e^*)$  and otherwise equals  $\Pi_{t+1} - \frac{1-\delta}{\delta}c(e^*)$ . The principal is indifferent between paying  $c(e^*)$  and

0 and so is willing to pay  $c(e^*)$  on the equilibrium path. If the agent deviates, however, the principal is equally willing to pay 0 in order to punish agent  $t$ . This indifference condition essentially uniquely pins down the possible continuation payoffs. Evidence helps cooperation by ensuring that, unlike Proposition 2, an agent cannot shirk and then promise the principal a high continuation payoff in exchange for paying less than the on-path transfer. For evidence to serve this role, however, the principal must be punished whenever an agent does not reveal evidence, since otherwise a shirking agent could promise non-disclosure to reward the principal for paying him.

A similar logic implies that non-disclosure triggers punishment for  $p < 1$  as well. In that case, principal-optimal equilibria are non-stationary, since a punishment phase is triggered whenever evidence is not available. This punishment phase also lowers the principal's payoff and implies that principal-optimal equilibrium effort will be lower than first-best even if players are very patient. The equilibrium described in Proposition 5 illustrates a particularly simple form of this punishment. In this equilibrium, decreasing  $p$  makes the principal worse off because it both increases the probability of the equilibrium transitioning to the frozen state and decreases equilibrium effort in the production phase.

Proposition 5 relies on an equilibrium construction that is fragile in two ways. First, the principal must be exactly indifferent between paying the equilibrium transfer and paying some smaller amount, since otherwise an agent could force the same transfer regardless of his effort level. Second, and more subtly, the principal's maximum expected continuation payoff must increase relatively rapidly in her transfer near  $s^*$ , since otherwise the agent could shirk and induce the principal to pay a transfer that is slightly smaller than  $s^*$ . The proof of Proposition 5 constructs an equilibrium that simultaneously satisfies these two conditions.

Our next result shows that a similar construction can be impossible if evidence is noisy. Formally, consider the *game with noisy evidence*, which is identical to the game with evidence and  $p = 1$  except that evidence is drawn from a non-degenerate distribution  $x \sim F(\cdot|s)$ . We give conditions on this distribution under which agents always shirk in equilibrium.

**Proposition 6.** *Consider the game with noisy evidence and  $\lambda = 1$ . Suppose that for any  $s \geq 0$ ,*

$$F(\cdot|s) = (1 - g(s))F_L(\cdot) + g(s)F_H(\cdot), \quad (1)$$

*where  $g(s) : \mathbb{R}_+ \rightarrow [0, 1]$  is strictly increasing, strictly concave, twice continuously differentiable, and satisfies  $g(0) = 0$  and  $\lim_{s \rightarrow 1} g(s) = 1$ . In any equilibrium of the game with noisy evidence,  $e_t = 0$  in each  $t \geq 0$  on the equilibrium path.*

**Proof:** See Appendix A.

To see the proof of Proposition 6, consider period  $t$  of an equilibrium. Let  $\underline{\pi}(s)$  and  $\bar{\pi}(s)$  be the principal's minimum and maximum expected payoffs if she pays  $s$ , including both this transfer and her continuation payoff. As discussed above, the principal must be indifferent between the equilibrium transfer  $s^*$  and some smaller transfer  $\hat{s}$ , which requires  $\bar{\pi}(s^*) = \underline{\pi}(\hat{s})$ . At the same time, agent  $t$  should not be able to force the principal to pay slightly less than  $s^*$ , which in particular requires  $\bar{\pi}(s^* - \epsilon) \leq \underline{\pi}(\hat{s})$  for small  $\epsilon > 0$ . These two conditions are impossible to satisfy simultaneously if  $\bar{\pi}(\cdot)$  is strictly concave, as it must be if  $F(\cdot|s)$  satisfies (1). In that case, any attempt to motivate an agent to exert effort simply creates an opportunity for that agent to shirk and earn rent, and so agents shirk in every equilibrium.

We can re-interpret the disclosure protocol as a restriction on the communication protocol, in the sense that it constrains the kinds of messages an agent can choose for each transfer. Since an agent can only reveal or conceal the actual payment, he cannot promise the principal a high continuation payoff unless she actually pays the amount she would on the equilibrium path, which means the principal has to incur a larger cost in order to give in to a hold-up attempt. Evidence limits the deviations available to agents and so decreases their power to extract surplus relative to the baseline model.

Propositions 5 and 6 highlight an inefficiency that is related to the moral hazard in teams problem from Holmstrom [1982]. Both settings suffer from a lack of what Fudenberg

et al. [1994] calls “pairwise identifiability:” given a public signal, players cannot tell whether that signal is the result of a deviation by the principal or a deviation by an agent. It is therefore not surprising that the principal’s maximum payoff is limited even if players are patient. While evidence can help support truthful communication, the resulting coordinated punishments must rely *entirely* on hard evidence, which means that evidence is of limited value to the principal unless it is always available.

## 4.2 Remedy 2: Bilateral Relationships

In the baseline game, the principal can refuse to pay an agent but cannot otherwise punish him. This section explores how future *bilateral* interactions between the principal and each agent can deter hold-up. We argue that these interactions do indeed facilitate coordinated punishments, but in a way that is limited by the strength of each bilateral relationship.

The **game with bilateral relationships** allows the principal to play a coordination game with each agent *after* that agent sends a message. At the end of each period in the baseline game, suppose that the principal and agent  $t$  simultaneously choose actions  $a_t^P \in \mathcal{A}^P$  and  $a_t^A \in \mathcal{A}^A$ , respectively, with  $a_t \equiv (a_t^P, a_t^A) \in \mathcal{A}$ . These actions are observed by the principal and agent  $t$  but not by any other agent. The principal’s and agent  $t$ ’s payoffs are  $\pi_t = e_t - s_t + v(a_t)$  and  $u_t = s_t - c(e_t) + w(a_t)$ , respectively.

The action space  $\mathcal{A}$  and payoffs  $(v(\cdot), w(\cdot))$  define a simultaneous move game, which we call the **bilateral relationship**. In equilibrium,  $a_t$  must correspond to a Nash equilibrium of the bilateral relationship. Let  $\mathcal{V}^*$  be the set of Nash equilibrium payoffs in the bilateral relationship. An equilibrium of the game with bilateral relationships essentially selects a payoff  $(v, w) \in \mathcal{V}^*$  at the end of each period  $t$ . This payoff can be used to reward or punish the principal or agent  $t$  for their earlier actions in the period. Note that  $\mathcal{V}^*$  is a convex because the players have access to a public randomization device.

Our next result characterizes the principal-optimal equilibrium in the game with bilateral relationships if the principal is patient.

**Proposition 7.** *Consider the game with bilateral relationships and suppose  $\lambda = 1$ . Define*

$$\begin{aligned}\bar{v} &= \max_{(v,w) \in \mathcal{V}^*} \{v\}, \\ \underline{v} &= \min_{(v,w) \in \mathcal{V}^*} \{v\}, \\ \underline{w} &= \min_{(v,w) \in \mathcal{V}^*} \{w\}, \\ \bar{L} &= \max_{(v,w) \in \mathcal{V}^*} \{v + w\}, \\ \underline{L} &= \underline{v} + \underline{w}.\end{aligned}$$

*Suppose  $\bar{L} > \underline{L}$ . There exists a  $\bar{\delta} < 1$  such that if  $\delta \geq \bar{\delta}$ , then  $e_t = e^*$  for each  $t \geq 0$  on the equilibrium path of any principal-optimal equilibrium, where*

$$c(e^*) = \min \{c(e^{FB}), (\bar{L} - \underline{L}) + (\bar{v} - \underline{v})\}. \quad (2)$$

*The principal's payoff is  $\Pi^* = e^* - c(e^*) + \bar{L} - \underline{w}$  and each agent earns  $\underline{w}$ .*

**Proof:** See Appendix A.

As in the baseline game, the proof of Proposition 7 identifies two incentive constraints that must be satisfied in any equilibrium. First, the principal must be willing to pay the equilibrium transfer rather than renege, where the punishment for renegeing potentially includes lower payoffs in both the bilateral relationship and future interactions with agents who learn of the deviation. Second, each agent must be willing to exert the equilibrium effort rather than shirking and holding up the principal. The bilateral relationship makes shirking less attractive by both punishing the agent for shirking and by punishing the principal for paying an agent who has shirked. We then construct an equilibrium that maximizes total surplus subject to these necessary conditions and then gives all of that surplus to the principal. The requirement that  $\delta \geq \bar{\delta}$  rules out a corner solution that arises if the principal's continuation payoff is very low, but it is otherwise not essential for the basic intuition.

The expression (2) highlights two ways in which bilateral relationships encourage effort

in equilibrium. First, as represented by  $(\bar{L} - \underline{L})$  in (2),  $a_t$  can directly reward or punish agent  $t$ 's effort choice or the principal's payment. In addition to this direct effect, which is familiar from the literature on relational contracting, bilateral relationships can also render coordinated punishments more effective by facilitating truthful communication among agents. This second effect is represented by  $(\bar{v} - \underline{v})$  and arises because  $a_t$  can be used to reward the principal to refusing to pay, or punish her for paying, an agent who has deviated. An agent consequently cannot demand the entire difference between the highest and lowest continuation payoff after he deviates, which means that bilateral relationships facilitate truthful communication and hence coordinated punishments. If these coordinated punishments are too severe, however, an agent can still hold up the principal. Equilibrium effort is therefore limited by the strength of the bilateral relationship.

Rather than modeling the bilateral relationship as an abstract simultaneous-move game, we could have alternatively considered a setting with long-lived agents who interact repeatedly with the principal. Appendix C takes this approach. We develop a repeated game with  $N$  long-lived agents and private monitoring. In each period, one agent is randomly chosen to play a favor-trading game with the principal. Agents do not observe one another's relationships but can commit to a communication protocol whenever they are chosen. While long-lived agents substantially complicate the analysis, we can show that (i) an agent's bilateral relationship with the principal facilitates coordinated punishments, and (ii) as in Proposition 7, the strength of that bilateral relationship tightly constrains how severe coordinated punishments can be in equilibrium. Thus, while some of the details change, our basic intuition extends to a setting with long-lived agents.

## 5 Interpreting the Communication Protocol

This section interprets the commitment assumption that lies at the heart of our analysis.

The communication protocol is powerful in our setting because it allows agents to credibly

signal which messages they will send following each transfer. In our baseline model, for instance, agents are indifferent among all possible messages. Without commitment, we can always construct an equilibrium in which an agent who shirks does not follow through on his threats and consequently cannot hold up the principal. Even in the game without commitment, however, there exist equilibria in which agents' hold-up threats are credible. Our next result shows that we can interpret our commitment assumption as an equilibrium refinement of the game without commitment.

**Proposition 8.** *In the baseline game, the game with evidence, or the game with bilateral punishments, for any equilibrium of the game with  $\lambda \in [0, 1]$ , there exists an equilibrium of the game with  $\lambda = 0$  which induces the same distribution over  $\{e_t, s_t\}_{t=0}^{\infty}$ .*

**Proof:** See Appendix A.

Proposition 8 follows immediately from Proposition 4 for the baseline game, and a nearly identical argument proves the result for the game with evidence. To prove the result for the game with bilateral relationships, we construct an equilibrium in which agents are punished for threatening hold-up regardless of whether or not they follow through on that threat. Agents' payoffs are then independent of their actual messages, so they are willing to follow the communication protocol.

Proposition 8 links our commitment assumption to the literature on equilibrium refinements, particularly refinements that capture a notion of bargaining in repeated games. For instance, our analysis provides a complementary view to Miller and Watson [2013]. We model bargaining in terms of a commitment assumption, an approach that disciplines our study of solutions to the hold-up problem in Section 4. Moreover, while Miller and Watson [2013] embed a notion of renegotiation-proofness in their model of bargaining, our analysis separates hold-up threats from renegotiation and so our equilibria may entail inefficient continuation play.

Punishments that rely on word-of-mouth are particularly susceptible to abuse in our set-

ting for two reasons. First, an agent who learns of a deviation through a message cannot disentangle two possibilities: either the principal truly deviated, or the communicator deviated by trying to hold up the principal. Therefore, that agent cannot tell whether the principal should actually be punished or not. Second, coordinated punishments are valuable only if an agent’s bilateral relationship with the principal is relatively weak. But then the principal has little recourse to directly punish an agent who engages in hold-up.

To illustrate this second point, recall that each agent is indifferent among messages in the baseline game. Consequently, if agents have even a mild preference for following through on their threats, they will act as if they are committed to those threats. To make this point, we consider a game with  $\varepsilon$ -**compliance preferences**, which is identical to the baseline game except that an agent does not commit to his communication protocol. Instead, immediately before each agent chooses his message, he observes his own preference for sending the message specified by his protocol. With probability  $\lambda \in [0, 1]$ , he earns an extra  $\varepsilon > 0$  if he chooses  $m_t = \mu_t(s_t)$ ; otherwise, his utility is independent of his message.

We show that the set of equilibrium outcomes in the game with  $\varepsilon$ -compliance preferences is identical to that of baseline game with commitment probability  $\lambda$ .

**Proposition 9.** *For any  $\varepsilon > 0$  and  $\lambda \in [0, 1]$ , there exists an equilibrium of the baseline game that induces a distribution over actions if and only if there exists an equilibrium of the corresponding game with  $\varepsilon$ -compliance preferences with that distribution over actions.*

**Proof:** See Appendix A.

Proposition 9 follows from the fact that even a mild preference for choosing  $m_t = \mu_t(s_t)$  makes this message the agent’s unique best response. Consequently, a given communication protocol generates an identical mapping from payments to messages regardless of whether an agent is committed or simply has  $\varepsilon$ -compliance preferences. This result extends to the game with evidence so long as agents also have a mild preference for following their disclosure protocols. It does not extend to the game with bilateral relationships, since a punishment

in the coordination game can overcome a mild intrinsic preference for following through on a threat. In that setting, Proposition 9 would require that agents' preferences for following the communication protocol are stronger than the *most severe* punishment in the bilateral relationship.

## 6 Discussion and Conclusion

In many settings, businesses and individuals cooperate with one another only because they fear that partners will spread word of any misbehavior. This paper proposes an under-explored downside to communication as a way to coordinate punishments. We argue that agents may abuse messages intended to report deviations to instead threaten the principal. These threats increase the agents' payoffs from shirking and so limit equilibrium cooperation. We also suggest that hard evidence and bilateral relationships are both imperfect solutions to this problem that suffer from their own downsides.

In our setting, the principal and agents play highly asymmetric roles and so rely on communication to very different extents. A natural next step would be to consider 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 hold one another up. How do players cooperate in the presence of two-sided hold-up? 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 both a useful building block and an informative first step towards analyzing these types of questions.

## References

- Nageeb Ali and Ce Liu. Conventions and coalitions in repeated games. Working Paper, 2018.
- Nageeb Ali and David Miller. Ostracism and forgiveness. *American Economic Review*, 106(8):2329–2348, 2016.
- Nageeb Ali, David Miller, and David Yang. Renegotiation-proof multilateral enforcement. 2016.
- Isaiah Andrews and Daniel Barron. The allocation of future business: Dynamic relational contracts with multiple agents. *American Economic Review*, 106(9):2742–2759, 2016.
- Chris Arnold and Robert Smith. Bad form, wells fargo. NPR, 10 2016. URL <https://www.npr.org/templates/transcript/transcript.php?storyId=499805238>.
- Yu Awaya and Vijay Krishna. Communication and cooperation in repeated games. Working Paper, 2018.
- George Baker, Robert Gibbons, and Kevin Murphy. Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, 109(4):1125–1156, 1994.
- George Baker, Robert Gibbons, and Kevin J. Murphy. Relational contracts and the theory of the firm. *The Quarterly Journal of Economics*, 117(1):39–84, 2002.
- Daniel Barron and Michael Powell. Policies in relational contracts. Working Paper, 2018.
- Daniel Barron, Jin Li, and Michal Zator. Productivity and debt in relational contracts. Working Paper, 2018.
- Nachman Ben-Yehuda. The european witch craze of the 14th to 17th centuries: a sociologist’s perspective. *The American Journal of Sociology*, 86(1):1–31, 1980.

- B. Douglas Bernheim and Michael Whinston. Incomplete contracts and strategic ambiguity. *The American Economic Review*, 88(4):902–932, 1998.
- Lisa Bernstein. Beyond relational contracts: Social capital and network governance in procurement contracts. *Journal of Legal Analysis*, 7(2):561–621, 2015.
- Simon Board. Relational contracts and the value of loyalty. *American Economic Review*, 101(7):3349–3367, 2011.
- Clive Bull. The existence of self-enforcing implicit contracts. *The Quarterly Journal of Economics*, 102(1):147–159, 1987.
- Olivier Compte. Communication in repeated games with imperfect private monitoring. *Econometrica*, 66(3):597–626, 1998.
- Yuk-Fai Fong and Jin Li. Relational contracts, limited liability, and employment dynamics. 2017.
- Richard Freeman. The exit-voice tradeoff in the labor market: Unionism, job tenure, quits, and separations. *The Quarterly Journal of Economics*, 94(4):643–673, 1980.
- Richard Freeman and Edward Lazear. An economic analysis of works councils. In Joel Rogers and Wolfgang Streeck, editors, *Works Councils: Consultation, Representation, and Cooperation in Industrial Relations*, pages 27–52. 1995.
- Richard Freeman and James Medoff. The two faces of unionism. *The Public Interest*, 57: 69–93, 1979.
- Drew Fudenberg, David Levine, and Eric Maskin. The folk theorem with imperfect public monitoring. *Econometrica*, 62(5):997–1039, 1994.
- Robert Gibbons and Rebecca Henderson. What do managers do? exploring persistent performance differences among seemingly similar enterprises. In Robert Gibbons and John Roberts, editors, *Handbook of Organizational Economics*, pages 680–731. 2013.

- Susanne Goldlucke and Sebastian Kranz. Reconciling relational contracting and hold-up: A model of repeated negotiations. Working Paper, 2017.
- Avner Greif, Paul Milgrom, and Barry Weingast. Coordination, commitment, and enforcement: The case of the merchant guild. *Journal of Political Economy*, 102(4):745–776, 1994.
- Yingni Guo and Johannes Hörner. Dynamic allocation without money. Working Paper, 2018.
- Yingni Guo and Eran Shmaya. Costly miscalibration. Working Paper, 2018.
- Marina Halac. Relational contracts and the value of relationships. *American Economic Review*, 102(2):750–779, 2012.
- Marina Halac. Investing in a relationship. *RAND Journal of Economics*, 46(1):165–186, 2015.
- Bengt Holmstrom. Moral hazard in teams. *The Bell Journal of Economics*, 13(2):324–340, 1982.
- Johannes Hörner and Nicolas Lambert. Motivational ratings. Working Paper, 2017.
- Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.
- Michihiro Kandori and Hitoshi Matsushima. Private observation, communication, and collusion. *Econometrica*, 66(3):627–652, 1998.
- Tobias Klein, Christian Lambertz, Giancarlo Spagnolo, and Konrad Stahl. Last minute feedback. Working Paper, 2006.
- Tobias Klein, Christian Lambertz, and Konrad Stahl. Market transparency, adverse selection, and moral hazard. *Journal of Political Economy*, 124(6):1677–1713, 2017.

- Jonathan Levin. Multilateral contracting and the employment relationship. *The Quarterly Journal of Economics*, 117(3):1075–1103, 2002.
- Jonathan Levin. Relational incentive contracts. *The American Economic Review*, 93(3): 835–857, 2003.
- Jin Li, Niko Matouschek, and Michael Powell. Power dynamics in organizations. *American Economic Journal: Microeconomics*, 2017. Forthcoming.
- Elliot Lipnowski and Joao Ramos. Repeated delegation. 2017.
- Elliot Lipnowski, Doron Ravid, and Denis Shishkin. Persuasion via weak institutions. Working Paper, 2018.
- Steffen Lippert and Giancarlo Spagnolo. Networks of relations and word-of-mouth communication. *Games and Economic Behavior*, 72(1):202–217, 2011.
- R.G. Lipsey and Kelvin Lancaster. The general theory of second best. *The Review of Economic Studies*, 24(1):11–32, 1956-1957.
- Bentley MacLeod and James Malcomson. Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica*, 57(2):447–480, 1989.
- James Malcomson. Relational incentive contracts. In Robert Gibbons and John Roberts, editors, *Handbook of Organizational Economics*, pages 1014–1065. 2013.
- James Malcomson. Relational contracts with private information. *Econometrica*, 84(1): 317–346, 2016.
- David Miller and Joel Watson. A theory of disagreement in repeated games with bargaining. *Econometrica*, 81(6):2303–2350, 2013.
- David Miller, Trond Olsen, and Joel Watson. Relational contracting, negotiation, and external enforcement. Working Paper, 2018.

- Daehong Min. Bayesian persuasion under partial commitment. Working Paper, 2017.
- Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK, 1990.
- Jason Proctor. Disgruntled bride ordered to pay 115k after defamatory posts ruin chinese wedding-photo business, 2018. URL <http://www.cbc.ca/news/canada/british-columbia/chinese-wedding-weibo-defamation-1.4556433>.
- Luis Rayo and Ilya Segal. Optimal information disclosure. *Journal of Political Economy*, 118(5):949–987, 2010.
- David Robinson and Toby Stuart. Network effects in the governance of strategic alliances. *Journal of Law, Economics, and Organizations*, 23(1):242–273, 2006.
- Joel Watson. Perfect bayesian equilibrium: General definitions and illustrations. 2016.
- Alexander Wolitzky. Cooperation with network monitoring. *The Review of Economic Studies*, 80(1):395–427, 2013.

# A Proofs

## A.1 Proof of Proposition 3

Each player's equilibrium payoff is at least 0: the principal can choose not to pay the agents, and agent  $t$  can choose not to exert effort or pay the principal. Let  $\bar{\Pi}$  be the highest equilibrium payoff for the principal. If  $\bar{\Pi} = 0$  then we are done, so suppose  $\bar{\Pi} > 0$  and consider the equilibrium profile that gives the principal  $\bar{\Pi}$ .

In  $t = 0$ ,  $e_0 > 0$ , since otherwise  $\bar{\Pi} \leq \delta \bar{\Pi}$  which leads to a contradiction. For each  $m \in M$ , let  $\Pi_1(m)$  be the principal's continuation payoff if  $m_0 = m$ . Let  $\bar{\Pi}_1 = \sup_m \Pi_1(m)$  and  $\underline{\Pi}_1 = \inf_m \Pi_1(m)$ , with corresponding messages  $\bar{m}$  and  $\underline{m}$ . Agent 0 is willing to choose  $e_0 > 0$  only if  $s_0 - c(e_0) \geq 0$ , so  $s_0 > 0$ . The principal is willing to pay  $s_0$  only if

$$s_0 \leq \frac{\delta}{1 - \delta} (\bar{\Pi}_1 - \underline{\Pi}_1), \quad (3)$$

so  $\bar{\Pi}_1 > \underline{\Pi}_1$ .

Consider the following deviation strategy by agent 0:  $e_0 = 0$  and

$$\mu_0(s) = \begin{cases} \underline{m} & s_0 < s^*, \\ \bar{m} & s_0 \geq s^*, \end{cases}$$

for some  $s^* \geq 0$ . Following this deviation, the principal's unique best response is to pay  $s^*$  if

$$(1 - \delta)(-s^*) + \delta (\lambda \bar{\Pi}_1 + (1 - \lambda) \underline{\Pi}_1) > \delta (\lambda \underline{\Pi}_1 + (1 - \lambda) \bar{\Pi}_1),$$

where the left-hand side of this inequality is the *smallest* principal payoff if she pays  $s^*$  and the right-hand side is the *largest* payoff if she does not. Therefore, the principal's unique

best response to this deviation is to play  $s^*$  so long as

$$s^* < (2\lambda - 1) \frac{\delta}{1 - \delta} (\bar{\Pi}_1 - \underline{\Pi}_1).$$

Consequently,

$$u_0 \equiv s_0 - c(e_0) \geq \max \left\{ 0, (2\lambda - 1) \frac{\delta}{1 - \delta} (\bar{\Pi}_1 - \underline{\Pi}_1) \right\} \quad (4)$$

in any equilibrium.

Combining (3) and (4) implies that a necessary condition for  $e_0$  and  $s_0$  is that

$$c(e_0) + \max \left\{ 0, (2\lambda - 1) \frac{\delta}{1 - \delta} (\bar{\Pi}_1 - \underline{\Pi}_1) \right\} \leq s_0 \leq \frac{\delta}{1 - \delta} (\bar{\Pi}_1 - \underline{\Pi}_1). \quad (5)$$

We claim that (5) is sufficient for  $e_0$  and  $s_0$  to be equilibrium choices in  $t = 0$ , given  $\Pi_1(\cdot)$ .

Indeed, consider the following strategy profile:

1. Agent 0 chooses  $e_0$  and  $\mu_0(s) = \begin{cases} \underline{m}, & \text{if } s_0 < c(e_0) + \max \left\{ 0, (2\lambda - 1) \frac{\delta}{1 - \delta} (\bar{\Pi}_1 - \underline{\Pi}_1) \right\}, \\ \bar{m}, & \text{otherwise.} \end{cases}$
2. The principal pays  $s_0$  satisfying (5) if agent 0 has not deviated, and some  $\tilde{s}_0 \leq \max \left\{ 0, (2\lambda - 1) \frac{\delta}{1 - \delta} (\bar{\Pi}_1 - \underline{\Pi}_1) \right\}$  that maximizes her continuation payoff otherwise.
3. If agent 0 can choose  $m_0$  (i.e., he is not committed), he chooses  $m_0 = \bar{m}$  if no deviation has occurred or if only agent 0 himself deviated, and  $m_0 = \underline{m}$  if the principal has deviated.

If agent 0 follows his equilibrium strategy, the principal has no profitable deviation because  $\frac{\delta}{1 - \delta} (\bar{\Pi}_1 - \underline{\Pi}_1) \geq s_0$ . If agent 0 deviates from  $(e_0, \mu_0)$ , then we need to show that the principal can best respond with some  $\tilde{s}_0 \leq \max \left\{ 0, (2\lambda - 1) \frac{\delta}{1 - \delta} (\bar{\Pi}_1 - \underline{\Pi}_1) \right\}$ . Indeed, the principal is willing to pay some  $\tilde{s}_0 > 0$  if

$$-\tilde{s}_0 + \frac{\delta}{1 - \delta} (\lambda \bar{\Pi}_1 + (1 - \lambda) \underline{\Pi}_1) \geq \frac{\delta}{1 - \delta} (\lambda \underline{\Pi}_1 + (1 - \lambda) \bar{\Pi}_1),$$

or equivalently,

$$\frac{\delta}{1-\delta}(2\lambda-1)(\bar{\Pi}_1 - \underline{\Pi}_1) \geq \tilde{s}_0.$$

The left-hand side is weakly less than agent 0's on-path payoff. So agent 0 has no profitable deviation from  $(e_0, \mu_0)$ . Agent 0 has no profitable deviation from  $m_0$  because he is indifferent among all messages. Continuation play is an equilibrium by assumption, so this strategy profile is an equilibrium, which proves that (5) is sufficient.

The principal-optimal equilibrium therefore solves

$$\begin{aligned} \bar{\Pi} &= \max_{e_0, s_0, \bar{\Pi}_1, \underline{\Pi}_1} && e_0 - s_0 + \frac{\delta}{1-\delta} \bar{\Pi}_1 \\ \text{s.t.} &&& (5) \quad \text{and} \quad 0 \leq \underline{\Pi}_1 \leq \bar{\Pi}_1 \leq \bar{\Pi}. \end{aligned} \tag{6}$$

If  $\lambda \leq \frac{1}{2}$ , then (5) simplifies to

$$c(e_0) \leq s_0 \leq \frac{\delta}{1-\delta} (\bar{\Pi}_1 - \underline{\Pi}_1).$$

Clearly,  $s_0 = c(e_0)$  maximizes the principal's payoff, in which case either  $e_0 = e^{FB}$  or  $c(e_0) = \frac{\delta}{1-\delta} (\bar{\Pi}_1 - \underline{\Pi}_1)$ . It is therefore optimal to set  $\bar{\Pi}_1 = \bar{\Pi}$ , which implies that play is stationary on-path, and (without loss)  $\underline{\Pi}_1 = 0$ . Consequently,  $\bar{\Pi} = e^{FB} - c(e^{FB})$  if  $c(e^{FB}) \geq \delta e^{FB}$ , and otherwise

$$c(e_0) = \frac{\delta}{1-\delta} \bar{\Pi} = \frac{\delta}{1-\delta} (e_0 - c(e_0))$$

or equivalently,  $e_0$  is the positive root of  $c(e_0) = \delta e_0$ , as desired.

Consider (6) for  $\lambda > \frac{1}{2}$ . The transfer  $s_0$  is optimally as small as possible, which by (5) implies that  $s_0 = c(e_0) + (2\lambda - 1) \frac{\delta}{1-\delta} (\bar{\Pi}_1 - \underline{\Pi}_1)$ . It is optimal to set  $\bar{\Pi}_1 = \bar{\Pi}$  because we can otherwise increase  $\bar{\Pi}_1, \underline{\Pi}_1$  holding  $\bar{\Pi}_1 - \underline{\Pi}_1$  constant. So  $e_t = e_0$  for all  $t \geq 0$  in any principal-optimal equilibrium.

Since  $s_0$  is increasing in  $\frac{\delta}{1-\delta} (\bar{\Pi}_1 - \underline{\Pi}_1)$ , it is optimal to set  $\underline{\Pi}_1$  so that both inequalities

in (5) hold with equality, which implies that  $s_0 = \frac{\delta}{1-\delta} (\bar{\Pi}_1 - \underline{\Pi}_1)$ . Therefore, we must have

$$\frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}_1) = c(e_0) + (2\lambda - 1) \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}_1)$$

or

$$2(1-\lambda) \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}_1) = c(e_0).$$

We can plug expressions for  $s_0$  and  $c(e_0)$  into (6) to rewrite this as an optimization with a single choice variable,  $\underline{\Pi}_1$ . That is,

$$\begin{aligned} \bar{\Pi} = \max_{\underline{\Pi}_1} \quad & c^{-1} \left( 2(1-\lambda) \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}_1) \right) + \frac{\delta}{1-\delta} \underline{\Pi}_1 \\ \text{s.t.} \quad & 0 \leq \underline{\Pi}_1 \leq \bar{\Pi} \end{aligned}$$

Since  $c^{-1}$  is concave and its derivative at 0 is  $+\infty$ , the optimal  $\underline{\Pi}_1$  is either interior or equal to 0. If the optimal  $\underline{\Pi}_1$  is interior, it is given by the first-order condition

$$(c^{-1})' \left( 2(1-\lambda) \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}_1) \right) = \frac{1}{2(1-\lambda)}.$$

In that case,  $e_0$  solves  $c'(e_0) = 2(1-\lambda) < 1$ . If  $\underline{\Pi}_1 = 0$ , then  $c(e_0) = (2-2\lambda) \frac{\delta}{1-\delta} \bar{\Pi}$ . This gives the equilibrium effort  $e^*(\lambda)$ .

In either case, we can then solve for  $\bar{\Pi}$ :

$$\bar{\Pi} = e_0 - \frac{c(e_0)}{2-2\lambda},$$

as desired. Since  $e - \frac{c(e)}{2-2\lambda}$  strictly increases in  $e$  for any  $e$  such that  $c'(e) \leq 2(1-\lambda)$ , and  $e^*(\lambda)$  strictly decreases in  $\lambda$ , the principal's payoff also strictly decreases in  $\lambda$ . ■

## A.2 Proof of Proposition 4

Let  $\lambda' > \lambda$ , and consider an equilibrium of the game with commitment probability  $\lambda'$ . We refer to this equilibrium as the *actual equilibrium*. Define the following strategy profile,

which we refer to as the *candidate profile*, of the game with commitment probability  $\lambda$ . In each history in each period  $t$ :

1. Agent  $t$  chooses  $e_t$  and  $\mu_t$  as in the actual equilibrium.
2. The principal chooses  $s_t$  as in the actual equilibrium.
3. For any  $\mu_t$ , agent  $t$  plays the following mixed strategy whenever he can freely choose  $m_t$ : with probability  $\frac{\lambda' - \lambda}{1 - \lambda}$ , he sends  $m_t = \mu_t(s_t)$ . Otherwise, he chooses  $m_t$  as in the actual equilibrium.

It suffices to show that players cannot profitably deviate in any period  $t \geq 0$ .

By construction, for any realized  $m_t$ , continuation payoffs from  $t + 1$  are identical in the candidate profile and the actual equilibrium. In stage (3), agent  $t$  is indifferent among messages and so is willing to follow the specified mixed strategy. Consequently, for any  $e_t$ ,  $\mu_t$ , the mapping from  $s_t$  to  $m_t$  is identical in the candidate profile and the actual equilibrium. The principal is therefore willing to pay the specified  $s_t$ . But then any choice of  $e_t$  and  $\mu_t$  induces the same  $s_t$  in the candidate profile and the actual equilibrium, so agent  $t$  is willing to play the specified  $e_t$  and  $\mu_t$ .

We conclude that the candidate profile is in fact an equilibrium of the game with commitment probability  $\lambda$  for any  $\lambda < \lambda'$ , as desired. ■

### A.3 Proof of Proposition 5

Let  $\bar{\Pi}$  be the highest equilibrium payoff for the principal. Let  $e^*$  and  $s^*$  be the equilibrium effort and transfer, respectively, in period 0 of the principal-optimal equilibrium. Let  $\bar{\Pi}_x$  and  $\underline{\Pi}_x$  be the highest and lowest equilibrium continuation payoffs, respectively, if the evidence realization is  $x$  for  $x \in \{\emptyset\} \cup \mathbb{R}_+$ .

Define

$$\hat{s} = \min \left\{ \arg \max_{s \in [0, s^*]} \left\{ -s + \frac{\delta}{1 - \delta} ((1 - p)\underline{\Pi}_\emptyset + p \min \{\underline{\Pi}_s, \underline{\Pi}_\emptyset\}) \right\} \right\}.$$

Intuitively,  $\hat{s}$  is the *smallest* transfer that maximizes the principal's payoff if her continuation payoff is as low as possible given her transfer.

The principal is willing to pay  $s^*$  only if

$$-s^* + \frac{\delta}{1-\delta} \left( (1-p)\bar{\Pi}_\emptyset + p \max \{ \bar{\Pi}_{s^*}, \bar{\Pi}_\emptyset \} \right) \geq -\hat{s} + \frac{\delta}{1-\delta} \left( (1-p)\underline{\Pi}_\emptyset + p \min \{ \underline{\Pi}_{\hat{s}}, \underline{\Pi}_\emptyset \} \right). \quad (7)$$

Agent 0 is willing to choose  $e^*$  only if

$$s^* - c(e^*) \geq \frac{\delta}{1-\delta} \left( \bar{\Pi}_\emptyset - \left( (1-p)\underline{\Pi}_\emptyset + p \min \{ \underline{\Pi}_{\hat{s}}, \underline{\Pi}_\emptyset \} \right) \right). \quad (8)$$

If this inequality does not hold, then agent 0 could deviate by choosing  $e_0 = 0$  and choosing the following communication protocol: if  $s_0 = s^* - c(e^*) + \varepsilon$  for a small  $\varepsilon > 0$ , send the message that leads to  $\bar{\Pi}_\emptyset$ ; otherwise, send the message that minimizes the principal's payoff. Faced with this communication protocol, the principal either pays  $s^* - c(e^*) + \varepsilon$  or  $\hat{s}$ . If (8) did not hold, then the principal would strictly prefer to pay  $s^* - c(e^*) + \varepsilon$  for sufficiently small  $\varepsilon > 0$ .

Let us suppose that (7) and (8) are sufficient conditions for equilibrium. In that case, it is without loss to set  $\hat{s} = 0$ . Indeed, if  $\hat{s} > 0$ , then

$$-\hat{s} + \frac{\delta}{1-\delta} \left( (1-p)\underline{\Pi}_\emptyset + p \min \{ \underline{\Pi}_{\hat{s}}, \underline{\Pi}_\emptyset \} \right) > \frac{\delta}{1-\delta} \left( (1-p)\underline{\Pi}_\emptyset + p \min \{ \underline{\Pi}_0, \underline{\Pi}_\emptyset \} \right), \quad (9)$$

which holds only if  $\underline{\Pi}_0 < \min \{ \underline{\Pi}_{\hat{s}}, \underline{\Pi}_\emptyset \}$ . We can therefore set  $\underline{\Pi}_0 \in [0, \min \{ \underline{\Pi}_{\hat{s}}, \underline{\Pi}_\emptyset \}]$  so that (9) holds with equality. This perturbation relaxes (8), does not affect (7), and results in  $\hat{s} = 0$ . So we can set  $\hat{s} = 0$ .

Given this perturbation, we can combine (7) and (8) into a single necessary condition:

$$\begin{aligned} \frac{\delta}{1-\delta} (p (\max \{\bar{\Pi}_{s^*}, \bar{\Pi}_\emptyset\} - \min \{\underline{\Pi}_0, \underline{\Pi}_\emptyset\}) + (1-p) (\bar{\Pi}_\emptyset - \underline{\Pi}_\emptyset)) &\geq s^* \\ &\geq c(e^*) + \frac{\delta}{1-\delta} (p (\bar{\Pi}_\emptyset - \min \{\underline{\Pi}_0, \underline{\Pi}_\emptyset\}) + (1-p) (\bar{\Pi}_\emptyset - \underline{\Pi}_\emptyset)). \end{aligned}$$

In a principal-optimal equilibrium,  $s^*$  is as small as possible, so

$$s^* = c(e^*) + \frac{\delta}{1-\delta} (p (\bar{\Pi}_\emptyset - \min \{\underline{\Pi}_0, \underline{\Pi}_\emptyset\}) + (1-p) (\bar{\Pi}_\emptyset - \underline{\Pi}_\emptyset)).$$

Our necessary condition can then be simplified to

$$\frac{c(e^*)}{p} \leq \frac{\delta}{1-\delta} (\max \{\bar{\Pi}_{s^*}, \bar{\Pi}_\emptyset\} - \bar{\Pi}_\emptyset). \quad (10)$$

Since neither  $\underline{\Pi}_0$  nor  $\underline{\Pi}_\emptyset$  appear in (10), setting  $\underline{\Pi}_0 \geq \underline{\Pi}_\emptyset = \bar{\Pi}_\emptyset$  minimizes  $s^*$ . Similarly, setting  $\bar{\Pi}_{s^*} = \bar{\Pi}$  is clearly optimal, since we can adjust  $\bar{\Pi}_\emptyset$  so that  $\bar{\Pi}_{s^*} - \bar{\Pi}_\emptyset$  is constant. Consequently, an upper bound on the principal's payoff is given by

$$\begin{aligned} \bar{\Pi} &= \max_{e^*, \bar{\Pi}_\emptyset} ((1-\delta)(e^* - c(e^*)) + \delta (p\bar{\Pi} + (1-p)\bar{\Pi}_\emptyset)) \\ \text{s.t.} & \quad (10) \quad \text{and} \quad 0 \leq \bar{\Pi}_\emptyset \leq \bar{\Pi}. \end{aligned} \quad (11)$$

If  $\bar{\Pi} > 0$ , then  $e^* > 0$ , in which case  $\bar{\Pi} > \bar{\Pi}_\emptyset$ . In that case, (10) holds with equality, since otherwise we could increase  $\bar{\Pi}_\emptyset$  and improve the principal's payoff. But then (10) implies that  $\frac{\delta}{1-\delta}\bar{\Pi}_\emptyset = \frac{\delta}{1-\delta}\bar{\Pi} - \frac{c(e^*)}{p}$ , which means that the objective of (11) simplifies to

$$\begin{aligned} \bar{\Pi} &= (1-\delta)(e^* - c(e^*)) + \delta\bar{\Pi} - (1-\delta)(1-p) \left( \frac{c(e^*)}{p} \right) \implies \\ \bar{\Pi} &= e^* - \frac{c(e^*)}{p}, \end{aligned}$$

as desired.

Let  $e^{max}(p) = \arg \max_e \left\{ e - \frac{c(e)}{p} \right\}$ . If  $\delta \geq \frac{c(e^{max}(p))}{pe^{max}(p)}$ , then  $e^* = e^{max}(p)$  and  $\bar{\Pi}_\emptyset$  solves (10)

with equality. Otherwise,  $\bar{\Pi}_\emptyset = 0$ , in which case  $e^*$  is the positive root of  $\delta = \frac{c(e)}{pe}$ .

It remains to show that the solution to this relaxed problem can be attained in equilibrium. We do so by showing that an equilibrium of the type described in the proposition attains this payoff. In particular, consider an equilibrium that starts in the “production phase.” In this phase, play in each period is:

1. Agent  $t$  chooses  $e_t = e^*$ ,  $\mu_t(s_t) = m$ , and  $\kappa_t(x_t) = x_t$ .
2. On-path, the principal pays  $s_t = c(e^*)$ . Following a deviation, the principal pays  $s_t = 0$ .
3. If  $\kappa_t(x_t) = c(e^*)$ , then stay in the production phase. Otherwise, stay in the production phase with probability  $\gamma \in [0, 1]$  and otherwise transition to the “frozen phase,” where  $\gamma$  solves  $\bar{\Pi}_\emptyset = \gamma\bar{\Pi}$  for the optimal  $\bar{\Pi}_\emptyset$  from (11).

In the frozen phase, which is absorbing, players play the one-shot equilibrium in each period and the principal earns 0.

Play in the frozen phase is clearly an equilibrium. In the production phase, the principal is willing to pay  $s_t = c(e^*)$  on the equilibrium path because

$$-c(e^*) + \frac{\delta}{1-\delta} (p\bar{\Pi} + (1-p)\gamma\bar{\Pi}) \geq \frac{\delta}{1-\delta} \gamma\bar{\Pi}$$

by choice of  $\gamma$  and by (10). Following any deviation by agent  $t$ , the principal pays 0. She is willing to pay 0 because (10) holds with equality. Given the principal’s actions, the agent can earn no more than 0 and so has no profitable deviation. So this strategy profile is an equilibrium that yields payoff

$$(1-\delta)(e^* - c(e^*)) + \delta p\bar{\Pi} + \delta(1-p)\bar{\Pi}_\emptyset = \bar{\Pi},$$

as desired. ■

## A.4 Proof of Proposition 6

Let  $f$ ,  $f_H$ , and  $f_L$  be the densities of  $F$ ,  $F_H$ , and  $F_L$ , respectively. Consider on-path play in period  $t \geq 0$  of an equilibrium. For each  $x \in \mathbb{R}$ , let  $\bar{P}i(x)$  and  $\underline{\Pi}(x)$  be the principal's maximum and minimum continuation payoffs if the evidence is  $x$ . Define

$$\Pi_E(s) = \int \bar{\Pi}(x)f(x|s)dx,$$

where

$$\Pi_E(\infty) = \int \bar{\Pi}(x)f_H(x|s)dx.$$

For any  $s \in \mathbb{R}_+$ , we can write

$$\begin{aligned} \Pi_E(s) &= \int \bar{\Pi}(x)(g(s)f_H(x) + (1 - g(s))f_L(x))dx \\ &= g(s) \int \bar{\Pi}(x)f_H(x)dx + (1 - g(s)) \int \bar{\Pi}(x)f_L(x)dx \\ &= g(s)\Pi_E(\infty) + (1 - g(s))\Pi_E(0). \end{aligned}$$

Define

$$\pi^D = \max_s \left\{ -s + \int_x \underline{\Pi}(x)f(x|s)dx \right\} \quad (12)$$

**Claim 1:** If  $e^* > 0$ , then  $\Pi_E(\infty) > \Pi_E(0)$ .

**Proof of Claim 1:** Suppose that  $e^* > 0$  but  $\Pi_E(\infty) \leq \Pi_E(0)$ . Note that  $s^* > 0$ . For small  $\epsilon > 0$ , consider the following deviation by agent  $t$ :  $e_t = 0$  and  $\mu_t, \kappa_t$  such that if  $s_t = s^* - \epsilon$ , the message and disclosure induce continuation payoff  $\bar{\Pi}(x_t)$  for each realization  $x_t$ , while for any other  $s_t$ , the message and disclosure induce continuation payoff  $\underline{\Pi}(x_t)$ . On the equilibrium path, the principal is willing to pay  $s^* > 0$  only if

$$-s^* + g(s^*)\Pi_E(\infty) + (1 - g(s^*))\Pi_E(0) \geq \pi^D. \quad (13)$$

Since  $g(\cdot)$  is strictly increasing,

$$-(s^* - \epsilon) + g(s^* - \epsilon)\Pi_E(\infty) + (1 - g(s^* - \epsilon))\Pi_E(0) > \pi^D. \quad (14)$$

Therefore, faced with this deviation, the principal has a strict incentive to pay  $s^* - \epsilon$ . We conclude that agent  $t$  can profitably deviate. Consequently,  $e^* = 0$  if  $\Pi_E(\infty) \leq \Pi_E(0)$ .

**Claim 2:** If  $e^* > 0$ , then  $\Pi_E(\cdot)$  is strictly concave.

**Proof of Claim 2:** Suppose  $e^* > 0$ . Then

$$\Pi_E'(s) = g'(s)(\Pi_E(\infty) - \Pi_E(0)) < 0,$$

since  $g'(\cdot) < 0$  and  $\Pi_E(\infty) > \Pi_E(0)$  by the previous claim.

**Completing the proof** Suppose  $e^* > 0$ . If (13) holds strictly, then (14) holds strictly for  $\epsilon > 0$  sufficiently small and so agent  $t$  has a profitable deviation. Therefore, (13) must hold with equality. Let  $\hat{s}$  be the smallest maximizer of (12). Then

$$-s^* + g(s^*)\Pi_E(\infty) + (1 - g(s^*))\Pi_E(0) \leq -\hat{s} + g(\hat{s})\Pi_E(\infty) + (1 - g(\hat{s}))\Pi_E(0).$$

We first claim that  $\hat{s} < s^*$ . If  $\hat{s} \geq s^*$ , then agent  $t$  can deviate in the way specified in Claim 1, but with  $\epsilon = 0$ . Since  $\hat{s}$  is the smallest transfer that maximizes (12), the principal pays no less than  $s^*$  following this deviation, where is therefore profitable.

To conclude, consider the difference

$$-s + \Pi_E(s) - (-s^* + \Pi_E(s^*)).$$

This difference is non-negative at both  $s = \hat{s}$  and  $s = s^*$ , where  $\hat{s} < s^*$ . It is strictly concave in  $s$  by Claim 2. We conclude that it must be strictly positive for  $s \in (\hat{s}, s^*)$ . But then

(13) implies that (14) holds for any  $\epsilon > 0$  such that  $s^* - \epsilon \in (\hat{s}, s^*)$ . Hence, agent  $t$  has a profitable deviation, so every equilibrium must entail  $e_t = 0$  in every  $t \geq 0$ . ■

## A.5 Proof of Proposition 7

Consider period  $t$  of an equilibrium in the game with bilateral relationships. Let on-path effort, transfer, and coordination game payoffs equal  $e^*$ ,  $s^*$ , and  $(v^*, w^*)$ , respectively. Let  $\bar{\Pi}$  and  $\underline{\Pi}$  be the highest and lowest continuation payoffs for the principal in this equilibrium, and let  $\bar{m}$  and  $\underline{m}$  be the messages that induce those payoffs.

The payment  $s^*$  must satisfy

$$s^* \leq \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}) + v^* - \underline{v}. \quad (15)$$

Agent  $t$  can always deviate by choosing  $e_t = 0$  and then choosing an optimal communication protocol. By deviating in this way, agent  $t$  earns a transfer of no less than  $\hat{s}$  in equilibrium, where  $\hat{s}$  is the largest transfer that satisfies

$$-s + \bar{v} + \frac{\delta}{1-\delta} \underline{\Pi} < -\hat{s} + \underline{v} + \frac{\delta}{1-\delta} \bar{\Pi}$$

for any  $s < \hat{s}$ . Therefore, agent  $t$  receives a transfer of at least

$$\hat{s} = \max \left\{ 0, \underline{v} - \bar{v} + \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}) \right\} - \varepsilon$$

for any  $\varepsilon > 0$  by deviating.

Given the argument above, a necessary condition for agent  $t$  to have no profitable deviation is

$$\max \left\{ 0, \underline{v} - \bar{v} + \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}) \right\} + \underline{w} \leq s^* - c(e^*) + w^*. \quad (16)$$

Combining (15) and (16), a necessary condition for equilibrium is that

$$\max \left\{ 0, \underline{v} - \bar{v} + \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}) \right\} + \underline{w} + c(e^*) - w^* \leq \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}) + v^* - \underline{v}. \quad (17)$$

Note that  $v^* + w^* \leq \bar{L}$ . If  $\underline{v} - \bar{v} + \frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi}) < 0$ , then increasing  $\frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi})$  relaxes (17). Suppose that we can increase  $\frac{\delta}{1-\delta} (\bar{\Pi} - \underline{\Pi})$  without bound. Then a necessary condition for (17) is that

$$\underline{v} - \bar{v} + \underline{w} + c(e^*) \leq \bar{L} - \underline{v},$$

or

$$c(e^*) \leq \bar{L} - \underline{L} + \bar{v} - \underline{v}. \quad (18)$$

Each agent  $t$  can earn no less than  $\underline{w}$ . The principal's payoff cannot exceed

$$\begin{aligned} \Pi^* &= \max_{e^*} \{ e^* - c(e^*) + \bar{L} - \underline{w} \} \\ \text{s.t.} & \quad (18). \end{aligned}$$

The solution to this problem is clearly the effort level given by (2). To prove the claim, therefore, it suffices to find a discount factor  $\bar{\delta} < 1$  such that for any  $\delta \geq \bar{\delta}$ , the principal earns  $\Pi^*$ . This will immediately imply that  $e_t$  satisfies (2) in each period on the equilibrium path.

We construct a strategy profile that attains the upper bound and give conditions under which it is an equilibrium. Let  $e^*$  satisfy (2) and  $(v^*, w^*) \in \arg \max_{(v,w) \in \mathcal{V}^*} \{v + w\}$ . Define

$$s^* = \underline{w} - w^* + c(e^*).$$

Let  $\rho$  be the solution to

$$\frac{\delta}{1-\delta} \rho (\Pi^* - \underline{v}) = \bar{v} - \underline{v}. \quad (19)$$

Since  $\bar{L} - \underline{L} > 0$  and  $\Pi^* \geq \bar{L} - \underline{w}$ ,  $\Pi^* - \underline{v} > 0$ . Consequently, there exists a  $\bar{\delta} < 1$  such

that  $\rho \in [0, 1]$  if  $\delta \geq \bar{\delta}$ .

For  $\delta \geq \bar{\delta}$ , consider the following strategy. In each  $t$ :

1. If  $m_{t'} = \bar{m}$  for all  $t' < t$ :

(a) Agent  $t$  chooses  $e_t = e^*$  to satisfy (2), and

$$\mu_t(s_t) = \begin{cases} \underline{m} & s_t < s^* \\ \bar{m} & s_t \geq s^* \end{cases}.$$

(b) If agent  $t$  did not deviate, the principal pays  $s_t = s^*$  if  $s^* > 0$ , or agent  $t$  pays  $s_t = s^*$  if  $s^* < 0$ . If agent  $t$  has deviated, the principal pays nothing and agent  $t$  pays  $s_t = -(w^P - \underline{w})$ , where  $(v^P, w^P) \in \mathcal{V}^*$  satisfies  $v^P = \bar{v}$ .

(c) If neither the principal nor agent  $t$  has deviated in period  $t$ ,  $(v_t, w_t) = (v^*, w^*)$  in the coordination game. If the principal has deviated in  $s_t$  (regardless of agent  $t$ 's actions),  $(v_t, w_t)$  satisfies  $v_t = \underline{v}$ . If agent  $t$  has deviated in  $e_t$  or  $\mu_t$  but pays  $s_t = -(w^P - \underline{w})$ , then  $(v_t, w_t) = (v^P, w^P)$ ; if he has deviated in  $s_t$ , then  $(v_t, w_t)$  satisfies  $w_t = \underline{w}$ .

2. If  $m_{t-1} \neq \bar{m}$ : Using the public randomization device, with probability  $(1 - \rho) \in [0, 1]$  continuation play is as if  $m_{t-1} = \bar{m}$ . Otherwise, play  $e_t = 0$  and  $(v_t, w_t)$  such that  $v_t = \underline{v}$  in this and all future periods.

Under this strategy profile, the principal earns  $\Pi^*$  on the equilibrium path. Therefore, it suffices to show that this strategy profile is an equilibrium.

The players have no profitable deviation from Step 2 of this equilibrium, since there exists an equilibrium of the coordination game with  $v_t = \underline{v}$ . Therefore, the principal's continuation payoff equals  $\underline{v}$  in this punishment phase.

In Step 1(c), all specified payoffs are elements of  $\mathcal{V}^*$  and so players cannot profitably deviate.

In Step 1(b), suppose agent  $t$  has not yet deviated. If  $s^* > 0$ , then the principal has no profitable deviation if

$$-s^* + v^* + \frac{\delta}{1-\delta}\Pi^* \geq \underline{v} + \frac{\delta}{1-\delta}((1-\rho)\Pi^* + \rho\underline{v})$$

or

$$-(\underline{w} - w^* + c(e^*)) + v^* - \underline{v} + \frac{\delta}{1-\delta}\rho(\Pi^* - \underline{v}) \geq 0.$$

By the definition of  $\rho$ , this inequality is equivalent to

$$-c(e^*) + \bar{L} - \underline{L} + \bar{v} - \underline{v} \geq 0,$$

which holds by (18). If  $s^* < 0$ , then agent  $t$  has no profitable deviation if

$$-s^* \leq w^* - \underline{w}$$

or

$$-(\underline{w} - w^* + c(e^*)) \leq w^* - \underline{w},$$

which holds because  $c(e^*) \geq 0$ .

Now, consider Step 1(b) and suppose agent  $t$  deviated in  $\mu_t$  or  $e_t$ . Agent  $t$  has no profitable deviation from  $s_t = -(w^P - \underline{w}) \leq 0$  because he earns  $w^P$  if he does so and  $\underline{w}$  otherwise. The principal pays nothing so long as for any  $\hat{s} > 0$ ,

$$-\hat{s} + \underline{v} + \frac{\delta}{1-\delta}\Pi^* \leq \bar{v} + \frac{\delta}{1-\delta}((1-\rho)\Pi^* + \rho\underline{v}).$$

Rearranging and setting  $\hat{s} = 0$ , we require

$$\frac{\delta}{1-\delta}\rho(\Pi^* - \underline{v}) \leq \bar{v} - \underline{v},$$

which is implied by (19). Therefore, the players have no profitable deviation in Step 1(b).

If agent  $t$  deviates in Step 1(a), then he earns no more than  $\underline{w}$  by the argument above. But he also earns  $\underline{w}$  on the equilibrium path. So agent  $t$  has no profitable deviation from Step 1(a). We conclude that the specified strategy profile is an equilibrium and therefore a principal-optimal equilibrium, and moreover, that in any principal-optimal equilibrium,  $e_t$  satisfies (2) in each period  $t \geq 0$  on the equilibrium path. ■

## A.6 Proof of Proposition 8

This result follows immediately from Proposition 4 in the baseline game. In the game with evidence, the agents are again indifferent over messages and so the result holds by an argument nearly identical to Proposition 4.

Consider the game with bilateral relationships. Let  $\sigma^*$  be an equilibrium for  $\lambda \geq 0$ , and consider the following strategy profile of the game with  $\lambda = 0$ : in each period  $t \geq 0$ ,

1. Agent  $t$  chooses  $e_t, \mu_t$  as in  $\sigma^*$ .
2. The principal chooses  $s_t$  as in  $\sigma^*$ .
3. Using the public randomization device: with probability  $\lambda$ , agent  $t$  chooses  $m_t = \mu_t(s_t)$ .  
Otherwise, agent  $t$  chooses  $m_t$  as in  $\sigma^*$ .
4. If the agent follows this message strategy,  $a_t$  is as in  $\sigma^*$ ; otherwise,  $a_t$  is chosen from those actions played in  $\sigma^*$  in order to minimize agent  $t$ 's payoff.

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  $e_t, \mu_t$ , or  $s_t$ , since continuation play is exactly as in  $\sigma^*$ . So this strategy profile is an equilibrium of the game without commitment, as desired. ■

## A.7 Proof of Proposition 9

Fix  $\varepsilon > 0$ . Consider an equilibrium of the baseline game with  $\varepsilon$ -compliance preferences. Agent  $t$  is indifferent among messages with probability  $1 - \lambda$  and so can send a message exactly as in the baseline game. With probability  $\lambda$ , agent  $t$ 's unique optimum is to choose  $m_t = \mu_t(s_t)$  and so he will do so in every equilibrium. So the mapping from  $\mu_t$  and  $s_t$  to  $m_t$  is identical to that in baseline game, which proves the result. ■