Discussion Paper Series – CRC TR 224

Discussion Paper No. 019
Project B 01

Sweet Lemons: Mitigating Collusion in Organizations

Colin von Negenborn*
Martin Pollrich**

February 2020
(First version: May 2018)

*Humboldt University Berlin, e-mail: von.negenborn@hu-berlin.de
**University of Bonn, e-mail: martin.pollrich@uni-bonn.de

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Sweet Lemons:

Mitigating Collusion in Organizations

Colin von Negenborn[a], Martin Pollrich[b]

February 13, 2020

Abstract

We show that mechanisms which generate endogenous asymmetric information fully mitigate collusion. In our model, an agent has private information and a supervisor observes a signal that is correlated with the agent's type. Agent and supervisor can form collusive side agreements. We study the implementation of social choice functions that condition on the agent's type and the supervisory signal. Our main result establishes that any social choice function that is implementable if the signal is public can also be implemented when the signal is private information and collusion is possible. Despite collusion, the signal is obtained for free, i.e., the supervisor does not receive an information rent. Our mechanism breaks collusion via endogenous creation of asymmetric information between agent and supervisor. The associated bargaining frictions prevent formation of collusive agreements, similar to the trade failure in the classical lemons market.

Keywords: Mechanism Design; Collusion; Correlation; Asymmetric Information; Random Incentives

JEL Codes: D82, D83, L51

[a]Humboldt University Berlin, e-mail: von.negenborn@hu-berlin.de

[b]University of Bonn, e-mail: martin.pollrich@uni-bonn.de

We would like to thank Helmut Bester, Françoise Forges, Vitali Gretschko, Daniel Krähmer, Benny Moldovanu, Anja Schöttner, Roland Strausz, Emanuele Tarantino and seminar participants at Aachen, Berlin, Bonn, Mannheim, Paris-Dauphine and Paris II for helpful comments and suggestions, as well as participants at Stony Brook Conference on Game Theory 2016, CTN Workshop 2017 Glasgow, CED 2017 York, EEA 2017 Lisbon, SING 2017 Paris, and SAET 2017 Faro. Colin von Negenborn is grateful for financial support by the German Research Foundation (DFG) through CRC TR 190 and by the Berlin Campus for Consumer Policies (BCCP). Martin Pollrich gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft—through CRC TR 224 (Project B01) and through Hausdorff Center of Mathematics—and from the DAAD.

1 Introduction

Asymmetric information is known to severely limit the allocation and distribution of resources. The frictions associated with information asymmetries give rise to new institutions whose major role lies in reducing the information gap. Examples range from certifiers asserting the (hidden) quality of products, to auditors verifying a company's financial statements.[1] The benefits these institutions provide crucially depend on their credibility, i.e., whether they reveal information truthfully. This paper considers collusion as a particular threat to credibility. Privately informed parties have an incentive to bribe information intermediaries into forging favorable information or concealing information altogether.[2]

We study a stylized setting with a single privately informed agent and a signal that is arbitrarily correlated with the agent's information. We compare two polar scenarios. Under *direct supervision* the signal is public, hence mechanisms can directly condition on its realization. Under *collusive supervision* the signal realization is only observed by the agent and a third party (the supervisor), who, in addition, can enter collusive side agreements. Our main result establishes that for any outcome which is implementable under direct supervision there is a mechanism implementing this outcome under collusive supervision with zero expected payment to the supervisor. Collusion has no bite: it neither impedes implementability, nor does it give rise to additional (collusive) information rents. This result holds for arbitrarily informative signals, comprising the special case of a fully informative signal.

We mitigate collusion by exploiting the above-mentioned frictions caused by asymmetric information.[3] Collusion is essentially a bilateral bargaining problem: the agent and the supervisor haggle over reports to be sent about the information each of them holds. Asymmetric information within the coalition shrinks the set of *feasible* outcomes for the coalition. The designer's problem

[1] Further examples include real estate brokers, schools rating the ability of their students, investment banks evaluating the quality of firms that want to raise capital, as well as (bond) rating agencies. Even the press (newspapers, scientific journals, etc.) serves as an information intermediary.

[2] A seller has an incentive to bribe the certifier into inflating a good's quality because high-valued goods sell at higher prices. A firm's manager has an incentive to bribe the auditor into forging financial statements when his wage depends on announced profits. The threat of collusion and corruption is documented both empirically and theoretically. According to IMF (2016) annual bribes exchanged worldwide exceed US$1 trillion. The theoretical literature on collusive supervision started with Tirole (1986), the literature review below provides a comprehensive overview.

[3] The literature on collusion noted early on that asymmetric information within potential coalitions weakens their detrimental effects. For early references see Laffont and Martimort (1997, 2000), Jeon and Menicucci (2005) and Che and Kim (2006).

is then to devise a mechanism in which the desired (non-collusive) outcome is the most attractive feasible outcome for the coalition. For our main result we devise a mechanism accomplishing this via shrinking the set of feasible outcomes to a singleton. Loosely speaking, collusion has no bite because the collusive parties cannot agree on a way of deviating.

The mechanism we devise creates endogenous asymmetric information. It randomizes the monetary payments to the agent and the supervisor and confidentially reveals the realization to the supervisor. By combining the design of these monetary payments and of the underlying randomness, the mechanism induces payoffs similar to those in the classical lemons market, c.f. Akerlof (1970).[4] Failure of trade is good news in our case, as it translates into a failure of collusion. Established approaches towards collusion use pre-existing informational asymmetries to make coalitional deviations unattractive. Our strategy for fighting collusion differs from that, by rendering collusion infeasible in the first place.

An implication of our results concerns the revelation principle. The literature acknowledges that there is no version of a revelation principle applicable to mechanism design with collusion. Most studies confine attention to deterministic mechanisms, where each party is asked to report its private information.[5] Our results imply that there is a strict gain from enriching the class of mechanisms: our grand mechanism creates endogenous asymmetric information by communicating with the players, before they communicate with the mechanism. For the special case of a fully informative signal, we show that a restriction to deterministic mechanisms severely restricts the set of implementable outcomes.

Another implication concerns the debate about organizational design, namely centralization vs. delegation. Previous literature shows that in the presence of collusion it may be without loss to delegate to the supervisor the task of contracting with the agent.[6] Typically collusion has some detrimental effect in these studies, thus delegation is strictly inferior: using our grand mechanisms in centralized contracting allows for mitigating those frictions.

[4]There is a slight difference regarding these payoffs. In the lemons market 'no trade' does not mean that no goods are traded, but only goods of the lowest quality. The payoffs in our mechanism are such that not even the lowest quality is traded because it would be inefficient to do so.

[5]Most studies on collusion restrict to mechanisms in which players simultaneously send messages. They prove a collusion-proofness principle (see, e.g., Laffont and Martimort (1997)) stating that w.l.o.g. message sets equal type spaces, and one can restrict attention to equilibria in which all reports are truthful and no collusion occurs. An exception is Chen and Micali (2012), who consider mechanisms that ask the agents also about the coalition they belong to. They show that such mechanisms allow for implementing the efficient allocation in an auction setting in dominant strategies.

[6]See for instance Faure-Grimaud et al. (2003) and Asseyer (2019).

The practical implementation of our mechanisms requires three aspects: randomization, conditioning payoffs on these random events, and confidential communication. Regarding randomization, the mechanisms we propose do not differ from stochastic mechanisms as analyzed in contract theory and mechanism design. To implement a stochastic mechanism the designer resorts to a randomization device or a third party. In an extension we argue that the designer's randomization is indeed credible, by devising a mechanism in which the expected payment following each draw is the same, and hence the designer is indifferent. Conditioning on random events is a standard feature of contracts and mechanisms. Remuneration typically consists of a fixed wage and a flexible bonus, the latter depends on random events such as stock prices, sales volume, etc. Our mechanism relies on confidential communication, which is already an essential feature of mechanism design with multiple agents (e.g., sealed bid auctions). We use this channel also in reverse direction, to allow the mechanism to communicate with its agents.

All three features mentioned above are present in the example of auditing.[7] Endogenous randomness is an inherent feature of the audit process, e.g., audits are triggered at random, auditors do not audit the entirety of a firm's accounts but only random samples, and there is random cross-checking. A manager's bonus depends on several indicators of performance, such as stock prices and sales volume. Many of these indicators are under scrutiny of auditors, hence the manager's wage indirectly depends on the auditor's report. Also, differential information between manager and auditor is commonplace (and designed). In a firm with several divisions, a particular division's manager may not know whether his or some other division is subject to an audit. Similarly, the auditee may not know whether the audit is part of a randomly triggered general investigation, or a targeted audit triggered by conspicuous data.[8]

We proceed as follows: after reviewing the literature, we introduce our model in Section 2. Section 3 presents our main result. Section 4 extends the baseline model into various dimensions: risk aversion, limited liability, ex-ante collusion participation constraints and commitment to devised mechanism. Section 5 concludes. All formal proofs are relegated to the Appendix.

[7]Another example combining the elements of our mechanism is that of "mystery shoppers". See `https://thenewinquiry.com/the-secret-shopper/` (accessed February 14, 2019). Hidden among regular customers, the mystery shopper secretly audits a company's staff. From the employee's perspective *any* customer could be an auditor. It is thus random (i) whether a customer reports, and (ii) what this customer reports on.

[8]The EU's *Horizon 2020* program prescribes "audits for periodical assessment of simplified cost forms", see pages 216–218 on `http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf` (accessed February 14, 2019). The entity currently being audited does not know whether the audit was triggered by inconsistencies in the provided data or whether it is just a periodical assessment.

Related Literature

We contribute to the literature on mechanism design with collusion, in particular collusion in hierarchies. Tirole (1986) introduces the principal–supervisor–agent hierarchy as the workhorse model for studying collusion in organizations. It consists of the privately informed agent, the supervisor who observes a signal correlated with the agent's type, and the principal as the mechanism designer. Tirole shows that supervision is beneficial despite collusion, but collusion gives rise to inefficiencies in form of productive distortions and additional information rents. The subsequent literature extends Tirole's model in various dimensions.[9] In contrast to our focus on implementability, this strand of literature typically focuses on maximizing the principal's profit, i.e., the cost of collusion in terms of revenue. In addition, to facilitate analytical results, these studies assume the agent's utility satisfy a single crossing property, moving the focus towards coalitional information rents and away from implementability. Our focus on implementation allows us to accommodate quite arbitrary preferences for the agent.

Kessler (2000), Burlando and Motta (2015) and Asseyer (2019) study models of hierarchies in which collusion can be completely mitigated.[10] Kessler (2000) and Burlando and Motta (2015) assume the supervisor's signal arrives only after accepting to participate in the grand mechanism, and the agent reports his private information before being able to collude with the supervisor. Further recall, that their results refer to the second-best revenue for the principal. Collusion still restricts implementability of some SCF. Asseyer (2019) assumes that the principal directly controls the supervisor's signal. He proves existence of a specific signal which allows the principal to realize the first-best revenue despite collusion and private information. Our result holds for *any* supervisory signal, and in particular, the first-best is implementable with a fully informative signal using our mechanisms.

Collusion has also been considered in general mechanism design settings.[11] Che and Kim

[9]Kofman and Lawarrée (1993), Kessler (2000) as well as Burlando and Motta (2015) keep Tirole's assumption of hard evidence, i.e., evidence can only be concealed and not forged. Baliga (1999), Faure-Grimaud et al. (2003), Celik (2009), Mookherjee et al. (2019) and Asseyer (2019) study collusion with soft supervisory information, where parties can claim any realized signal. Also, this literature differs with respect to the supervisor's preferences: Baliga (1999), Burlando and Motta (2015) as well as Asseyer (2019) assume limited liability, Faure-Grimaud et al. (2003) assume risk aversion, and all others assume risk neutrality. Kofman and Lawarrée (1993) assume risk neutrality, but limit the size of penalties.

[10]Faure-Grimaud et al. (2003) also point out a case where collusion is not detrimental: the special case of their model with a risk-neutral supervisor.

[11]For an early reference see Green and Laffont (1979) who show that Vickrey–Clarke–Groves mechanisms are vulnerable to coalition formation. See also Crémer (1996) and Chen and Micali (2012).

(2006) show that when agents' private information is independent, collusion causes no harm: any revenue the mechanism designer attains absent collusion can also be attained in the presence of collusion.[12] However, their results do not apply for (sub–) coalitions of size two when the agents' information is correlated, as is the case in our setting and most of the literature on collusion in hierarchies. In addition, and as pointed out earlier, our general approach towards fighting collusion differs, by rendering collusion infeasible in the first place.

This paper is also related to the literature on endogenous asymmetric information in contracting. In a related paper, Ortner and Chassang (2018) study the problem of deterring criminal activities via monitoring. An agent decides whether to commit a crime, and a monitor observes evidence of the crime. The agent may bribe the monitor into withholding evidence. They show that randomizing the monitor's wage and telling only the monitor her actual wage strictly reduces monitoring costs. Their setting differs from ours in two crucial aspects:[13] it involves moral hazard (the monitor observes the agent's *action*, not a signal correlated with his type), and the agent's payoff is exogenous. In contrast, we study a full-fledged mechanism design setting where the designer controls the payoffs of both the agent and the supervisor, allowing us to identify the exact channel through which endogenous asymmetric information is helpful. Most importantly, we prove that collusion can be *completely* mitigated.

Mechanisms that create endogenous asymmetric information have been previously investigated in moral hazard problems. In a team problem with budget–balanced payments Rahman and Obara (2010) show that mediated contracts mitigate moral hazard. Similar to our mechanism, their mediator sends confidential recommendations to the agents and transfers are conditioned on these recommendations and on the realized output.[14] Rahman (2012) uses similar mechanism to provide a monitor with incentives to actually monitor the hidden action of an agent.[15] While this literature uses endogenous private information to relax obedience constraints, in our setting (which is of pure adverse selection) private information increase bargaining frictions and thereby prevents collusion.

[12]Their results generalize earlier findings by Laffont and Martimort (1997) and Jeon and Menicucci (2005) for the case of independent private information.

[13]In addition, the optimality of random incentives in Ortner and Chassang (2018) depends on pre-existing patterns of private information, which is not the case in our setting.

[14]Strausz (2012) relates these *mediated contracts* back to the revelation principle.

[15]The mechanism in Rahman (2012) is prone to collusion. Pairing it with our methodology (adapted to moral hazard) may yield a collusion-proof mechanism that provides the monitor with incentives to monitor.

2 Model

The basic setting. There is a mechanism designer and a single agent. The mechanism designer controls the decision over an alternative $x \in \mathcal{X}$ and a monetary transfer $t \in \mathbb{R}$. The agent's utility when selecting alternative x and paying transfer t is $U_A(x,t,\theta) = u(x,\theta) + t$. The parameter $\theta \in \Theta = \{\theta_1, \ldots, \theta_n\}$ denotes the agent's type. We shall make no restrictions on the function $u(\cdot,\cdot)$, other than it being real-valued.[16] The set $\mathcal{X}$ of alternatives is arbitrary as well. Additive separability and risk-neutrality with respect to monetary payments to the agent are crucial assumptions, as will become clear from our analysis.

In addition to the agent's type θ there is a second piece of information: the signal $\tau \in \mathcal{T} = \{\tau_1, \ldots, \tau_m\}$. This signal is payoff-irrelevant (for the agent): it does not directly enter his utility. We allow for arbitrary correlation between θ and τ. Let $\pi = \left(\pi_{ij}\right)$ denote the prior, where $\pi_{ij} = \Pr(\theta = \theta_i, \tau = \tau_j)$ for all $1 \leq i \leq n$ and $1 \leq j \leq m$. We assume there are no redundant types or signals: $\sum_j \pi_{ij} \neq 0 \neq \sum_i \pi_{ij}$ for all i, j.[17] From the unconditional probabilities we can define the conditional probability $\pi_i^j = Pr(\theta_i|\tau_j) = \pi_{ij}/(\sum_i \pi_{ij})$.

We are interested in implementing social choice functions (SCF)

$$(\mathbf{x}, \mathbf{t}) \ : \ \Theta \times \mathcal{T} \to \mathcal{X} \times \mathbb{R}, \tag{1}$$

mapping the agent's type θ and the signal τ into a decision $\mathbf{x}(\theta,\tau) \in \mathcal{X}$ and a transfer to the agent $\mathbf{t}(\theta,\tau) \in \mathbb{R}$. Note that we explicitly allow the SCF to depend on the signal τ, though the latter is not payoff-relevant for the agent. First, we do not rule out payoff-relevance for outsiders, as this does not conflict with the implementation problem itself. Second, even if a SCF conditions only on the agent's type θ, its implementability may well be affected by the realization of τ (e.g., the case where τ partitions the agent's type space Θ).

[16] Most studies on collusion in hierarchies assume $u(x,\theta) = -x\theta$, e.g., Baliga (1999), Faure-Grimaud et al. (2003), Celik (2009), Asseyer (2019), Mookherjee et al. (2019). Others use a model with moral hazard and adverse selection, e.g., Tirole (1986), Kofman and Lawarrée (1993), Kessler (2000). The agent's cost of effort is $\psi(e)$ and there is one–to–one link between (equilibrium) effort and type. Both approaches implicitly assume a single-crossing condition, that facilitates implementability.

[17] This specification allows for a fully informative signal, where $p_{ii} = 1$ for all i. It also allows for signals that partition the type space, as in Celik (2009). Most studies use a binary type space and binary signal, e.g., Tirole (1986), Kofman and Lawarrée (1993), Baliga (1999), Kessler (2000), Faure-Grimaud et al. (2003).

Information and supervision. Throughout the analysis we assume that only the agent observes θ. Regarding the signal τ we distinguish two scenarios.

- *Direct supervision:* the signal τ is public.

- *Collusive supervision:* both a supervisor and the agent observe τ, but not the mechanism designer. In addition, agent and supervisor can collude, as specified below. The supervisor's utility only depends on her wage $w \in \mathbb{R}$, i.e., $U_S(x, w, \theta, \tau) = w$. The supervisor has no intrinsic motivation to misreport information, in particular the signal τ is also irrelevant for her payoff. In our baseline model we focus on a risk-neutral supervisor. Section 4.4 extends our results to the cases of limited liability and risk aversion.[18] Under collusive supervision we face a nested information structure: the agent observes both θ and τ, the supervisor observes only τ and any outsider (including the mechanism designer) observes neither θ nor τ.[19]

Mechanisms. The mechanism designer devises a *grand mechanism* (as opposed to the side mechanisms introduced next), which is an extensive form Bayesian game. Formally, a grand mechanism $\Gamma = \Big((\Upsilon_A, \Upsilon_S, \mu), \Sigma_A, \Sigma_S, g\Big)$ consists of type spaces Υ_A, Υ_S for agent and supervisor, a 'prior' $\mu \in \Delta(\Upsilon_A \times \Upsilon_S)$, strategy sets Σ_A, Σ_S for the agent and the supervisor, and an outcome function $g : \Upsilon_A \times \Upsilon_S \times \Sigma_A \times \Sigma_S \longrightarrow \mathcal{X} \times \mathbb{R} \times \mathbb{R}$, specifying an alternative and payments to the players. The mechanism first draws types (υ_A, υ_S) according to the distribution μ and confidentially reveals υ_i to player i $(i = A, S)$. Next, players choose their strategies from the sets Σ_A, Σ_S, and payoffs realize. For the special case of a deterministic mechanism we have $|\Upsilon_A| = |\Upsilon_S| = 1$, and for a standard stochastic mechanism we have $\Upsilon_A = \Upsilon_S$, and $\mu(\upsilon_A, \upsilon_S) = 0$ whenever $\upsilon_A \neq \upsilon_S$. But our definition of a grand mechanism also allows for endogenous asymmetric information.[20] Our analysis builds on a grand mechanism in which $|\Upsilon_A| = 1$ and $|\Upsilon_S| = m$. In such a mechanism the supervisor has (endogenous) private information vis–à–vis

[18] Tirole (1986), Baliga (1999), Kessler (2000), Celik (2009), Mookherjee et al. (2019) also study risk-neutral supervisors, Faure-Grimaud et al. (2003) study risk-averse supervisors, and Kofman and Lawarrée (1993), Asseyer (2019) study versions of limited liability.

[19] The agent has private information vis-à-vis the supervisor, and both the supervisor and the agent have private information vis-à-vis the mechanism designer.

[20] Nothing in the previous literature rules out this possibility. The mechanism designer himself is part of the mechanism, and can devise a game where he has a first move before agent and supervisor move. Imperfect and asymmetric observation of the designers initial move creates the desired information structure. With the usual assumptions on commitment, the designer commits to his initial move thus creating a Bayesian game.

the agent. We say the grand mechanism Γ implements SCF $(\mathbf{x}, \mathbf{t})$ if it exhibits an equilibrium (BNE) such that the resulting equilibrium allocation coincides with $(\mathbf{x}, \mathbf{t})$.

Collusion. We assume the agent and supervisor can coordinate their play of the grand mechanism. They can devise an enforceable[21] *side mechanism* committing them to strategy choices in the *grand mechanism* and to exchange side payments. We follow Laffont and Martimort (1997) in assuming that a disinterested third party proposes a collusive side mechanism Γ^c.[22]

The signal τ is symmetrically known to agent and supervisor, hence any side mechanism directly conditions on it. Players have two sources of private information at the collusion stage: exogenous and endogenous private information. The agent knows his type θ, which is not known to the supervisor and represents exogenous private information. The supervisor does not have any exogenous private information. In addition, agent and supervisor have endogenous private information from the grand mechanism, given by the realizations (υ_A, υ_S). We assume that this information is not verifiable at the side-contracting stage. A direct side mechanism $\Gamma^c = \left(\Theta \times \Upsilon_A, \Upsilon_S, \varsigma\right)$ asks each player to report their private information, and the outcome function $\varsigma : \Theta \times \Upsilon_A \times \Upsilon_S \to \Delta(\Sigma_A \times \Sigma_S) \times \mathbb{R} \times \mathbb{R}$ selects a vector of strategies to be played in the grand mechanism (potentially at random) and monetary side payments.

Invoking a revelation principle we can without loss of generality focus on direct side mechanisms in which both the agent and the supervisor reveal their private information truthfully. Denote $U^k(\theta, \upsilon_A, \upsilon_S)$ the expected utility of player $k = A, S$ in the grand mechanism, following reports $(\theta, \upsilon_A, \upsilon_S)$ to the side mechanism. Similarly, define $b^k(\theta, \upsilon_A, \upsilon_S)$ the associated side payment, and let $\bar{U}^k(\theta, \upsilon_A, \upsilon_S)$ denote the utility from playing the grand mechanism non-cooperatively. Each player reports his information truthfully in Γ^c if for all $\theta, \theta' \in \Theta, \upsilon_A, \upsilon'_A \in \Upsilon_A, \upsilon_S, \upsilon'_S \in \Upsilon_S$

$$\begin{aligned} &\mathbb{E}_{\upsilon_S}\Big[U^A(\theta, \upsilon_A, \upsilon_S) + b^A(\theta, \upsilon_A, \upsilon_S)] \geq \mathbb{E}_{\upsilon_S}\Big[U^A(\theta', \upsilon'_A, \upsilon_S) + b^A(\theta', \upsilon'_A, \upsilon_S)], \\ &\mathbb{E}_{\theta, \upsilon_A}\Big[U^S(\theta, \upsilon_A, \upsilon_S) + b^S(\theta, \upsilon_A, \upsilon_S)] \geq \mathbb{E}_{\theta, \upsilon_A}\Big[U^S(\theta, \upsilon_A, \upsilon'_S) + b^S(\theta, \upsilon_A, \upsilon'_S)]. \end{aligned} \tag{IC}$$

[21] Enforceability gives collusion its best chance. It is a short cut to capture in a static context the reputations of the third party, the agent and the supervisor which guarantee that the self-enforceability of these contracts would emerge in a repeated relationship.

[22] This modeling avoids signaling issues when the informed party proposes a side mechanism, as discussed in Laffont and Martimort (1997). In particular, our modeling gives collusion its best chance. Restricting to specific bargaining procedures would leave open the question as to whether our results depend on our specific modeling of collusive side bargaining.

Both players participate in Γ^c, if for all $\theta \in \Theta, \upsilon_A \in \Upsilon_A$, and $\upsilon_S \in \Upsilon_S$

$$\begin{aligned}
&\mathbb{E}_{\upsilon_S}\Big[U^A(\theta, \upsilon_A, \upsilon_S) + b^A(\theta, \upsilon_A, \upsilon_S)] \geq \mathbb{E}_{\upsilon_S}\Big[\bar{U}^A(\theta, \upsilon_A, \upsilon_S)], \\
&\mathbb{E}_{\theta,\upsilon_A}\Big[U^S(\theta, \upsilon_A, \upsilon_S) + b^S(\theta, \upsilon_A, \upsilon_S)] \geq \mathbb{E}_{\theta,\upsilon_A}\Big[\bar{U}^S(\theta, \upsilon_A, \upsilon_S)].
\end{aligned} \tag{IR}$$

Lastly, we assume that the third party itself does not possess any money. Formally, we require the side mechanism to be ex-ante budget balanced:

$$\mathbb{E}_{\theta,\upsilon_A,\upsilon_S}\Big[b^A(\theta, \upsilon_a, \upsilon_s) + b^S(\theta, \upsilon_a, \upsilon_s)\Big] \leq 0. \tag{BB}$$

With the weak notion of ex-ante budget-balancedness (as opposed to ex-post) we give collusion its best chance.[23] To summarize, we say a side-mechanism is feasible if it satisfies (IC), (IR) and (BB). A social choice function $(\mathbf{x}, \mathbf{t})$ is implementable under collusive supervision, if there is a grand mechanism that implements it (in the notion introduced above) and there is no feasible side mechanism that breaks equilibrium play.[24]

A crucial assumption for our results to hold is, that the side mechanism cannot condition on the realized payments (resp. utility levels) in the grand mechanism. In other words, we assume that all side payments are made *before* reporting to the grand mechanism, hence they cannot condition on the resulting allocation.[25] This is a reasonable assumption in many environments, e.g., where the decision at hand represents but one factor for the agent's (resp. the supervisor's) total remuneration. For instance, the agent may work on several tasks for the same principal (in our setting the designer), and receive a total wage bill without task-specific breakdown. Similarly, incentives may be provided not with money, but indirectly, e.g., via career concerns (promotions) or continuation payoffs (future job assignments). Typically, these choices depend on other factors as well, ruling out an exact quantification of the impact of the decision at hand.

Timing. To conclude this section we summarize the timing.

0. The mechanism designer selects a grand mechanism Γ.

[23] Though in many environments the two notions are equivalent, see Börgers and Norman (2009) for a discussion.

[24] Our focus does not lie on unique implementation. We seek to find *an* equilibrium that collusion cannot break. For most cases we cannot rule out equilibrium multiplicity for the grand mechanism. In particular there may be other equilibria involving collusion on the equilibrium path.

[25] Mezzetti (2004) shows that efficiency is implementable in a setting with interdependent values when the mechanism can condition on ex-post utilities. His result, applied to our setting, would render side-bargaining efficient and lead to collusion breaking the non-cooperative equilibrium.

1. Nature draws (θ, τ) from the prior π and informs the agent about θ, and both the supervisor and the agent about τ. In addition, nature draws (υ_A, υ_S) from μ and confidentially reveals them to the respective player.

2. A third party offers side contract Γ^c, and players simultaneously decide whether to participate. If at least one player rejects, we move directly to stage 4, otherwise to stage 3.

3. Players submit their reports to the side mechanism, determining a joint report and the exchange of side payments.

4. Both the agent and the supervisor select their strategies in the grand mechanism (in case of an effective side-contract the strategies determined therein), and the allocation is determined.

3 Analysis

3.1 Benchmark Cases

We begin our analysis with the discussion of two benchmark cases: direct supervision and non-collusive supervision.

Direct supervision. Under direct supervision the signal τ is publicly observable. A mechanism can directly condition on the realized τ.[26] The following lemma characterizes the set of implementable social choice functions.

Lemma 1. *A social choice function* $(\mathbf{x}, \mathbf{t})$ *is implementable under direct supervision if and only if*

$$u(\mathbf{x}(\theta, \tau), \theta) + \mathbf{t}(\theta, \tau) \geq u(\mathbf{x}(\theta', \tau), \theta) + \mathbf{t}(\theta', \tau) \quad \forall \theta, \theta' \in \Theta(\tau) \ \forall \tau \in \mathcal{T}, \tag{2}$$

where $\Theta(\tau) \coloneqq \{\theta \in \Theta | \Pr(\theta, \tau) > 0\}$.[27]

[26] In this case, there is no need for third-party supervision because the supervisor cannot provide additional information. Hence, we focus on mechanisms that elicit the agent's (residual) private information from his knowledge of θ, but ignore the supervisor.

[27] A direct mechanism allows the agent to report only types that have strictly positive probability under the realized signal, hence the set of reports for signal τ is $\Theta(\tau)$.

Note the special case of a *fully informative signal*, where $\mathcal{T} = \{\tau_1, \ldots, \tau_n\}$ and $\pi_{ii} = 1$ for all $1 \leq i \leq n$. In this case, $\Theta(\tau_i) = \{\theta_i\}$ for all i, and thus condition (2) has no bite. Consequently *any* social choice function $(\mathbf{x}, \mathbf{t})$ is implementable under direct supervision.

Non-collusive supervision. As a second benchmark we briefly study *non-collusive supervision*: the signal τ is not publicly known, but only observed by both the agent and the supervisor. However, for this benchmark we rule out collusion. To implement a SCF that was implementable under direct supervision we exploit the symmetric information about the signal. One possibility is to ask both the agent and the supervisor for reporting the signal and to penalize both in case their reports disagree.[28] Sufficiently large penalties avert *unilateral* deviations from truth-telling. Once the mechanism elicits the signal, the agent's truthful report of his type is implied by implementability under direct supervision.

Remark 1. *Any social choice function* $(\mathbf{x}, \mathbf{t})$ *that is implementable under direct supervision is also implementable under non-collusive supervision. Implementation does not require payments to the supervisor in equilibrium.*

3.2 Breaking Collusion

We now move on to study the general setting with collusion. Now, the mechanism has to deter *unilateral* as well as *joint* deviations. With respect to the former effective deterrence uses the fact that agent and supervisor have symmetric information on the signal τ. Any unilateral deviation from truthful reporting yields non-conforming reports.[29] Because such reports can only occur off the equilibrium path effective punishment is possible.

In contrast, joint deviations typically involve equilibrium reports and can neither be unambiguously detected nor effectively penalized.[30] The logic of deterrence presumes that the respective deviations are actually feasible. If there were no feasible deviations, the mechanism would not need to provide deterrent incentives. Our main result takes up that very point

[28] There are even simpler mechanisms available, but the idea of penalizing non-conforming reports will reappear in our grand mechanisms under collusion.

[29] It is known that there was a deviation, but not who deviated. Both players are penalized, which is irrelevant, though, because deviations occur off the equilibrium path.

[30] Che and Kim (2006) point out that a direct mechanism lacks instruments to simultaneously deter all unilateral and all group deviations. Their approach embraces group deviations by 'selling' the firm to the agents. With only two agents and correlated information "the transfer rule does not give a sufficient number of degrees of freedom to 'sell the firm to the agents' while preserving the original incentive design of t." (p. 1086)

and establishes that collusion does not restrict implementability. We devise a mechanism that renders all joint deviations from truthful reporting infeasible for the coalition. More precisely, we devise a mechanism that generates payoffs such that for any given (true) signal the agent and the supervisor cannot find a feasible side mechanism committing them to non-truthful reporting.

Proposition 1. *Any social choice function* $(\mathbf{x}, \mathbf{t})$ *that is implementable under direct supervision is also implementable under collusive supervision with zero expected payments to the supervisor.*

Our proof is constructive. We devise a mechanism with endogenous private information that implements the SCF $(\mathbf{x}, \mathbf{t})$. Regarding collusion, we prove that if agent and supervisor play the desired equilibrium there is only a single side mechanism that is feasible, which corresponds to the null side mechanism: it prescribes equilibrium play without side payments. In the following we describe the mechanism and sketch why side-contracting fails. In the appendix we provide a self-contained formal proof of Proposition 1.

For a given SCF $(\mathbf{x}, \mathbf{t})$ consider the following mechanism in form of a Bayesian game. First, nature draws a number $\kappa \in \{1, \ldots, m\}$ uniformly at random and confidentially reveals it to the supervisor. We will refer to κ as the 'bonus state', for reasons which will become clear shortly. Second, there is a reporting stage. The agent reports a signal (τ^a) and a type (θ^a), and the supervisor reports a signal (τ^s). Based on the reports $(\theta^a, \tau^a, \tau^s)$ and the bonus state κ the mechanism selects an alternative and payments as given in Table 1. The mechanism pays the supervisor a base wage ω. If both agent and supervisor report the bonus state τ_κ the supervisor receives a bonus β. Recall that only the supervisor knows the bonus state, i.e., which report triggers the bonus β. Furthermore, the mechanism selects the alternative and the agent's transfer according to $(\mathbf{x}, \mathbf{t})$. The agent receives in addition a lump-sum payment t, but his transfer reduces by γ in case the signal τ_κ is reported. Finally, non-conforming reports about the signal are penalized as in the shoot–the–liar mechanism introduced in Remark 1. The values ω and t are chosen such that in expectation the supervisor's wage is zero and the agent's transfer in state (θ, τ) equals $\mathbf{t}(\theta, \tau)$. It is then straightforward to verify the grand mechanism exhibits a non-cooperative equilibrium in which both agent and supervisor report truthfully, and this equilibrium implements $(\mathbf{x}, \mathbf{t})$.

The particular feature of our grand mechanism is the private information of the supervisor. Only he knows the bonus state, the agent does not. Independent of the true info state (θ, τ)

	$\tau^a = \tau^s = \tau_\kappa$	$\tau^a = \tau^s \neq \tau_\kappa$	$\tau^a \neq \tau^s$
$t(\theta^a, \tau^a, \tau^s, \kappa)$	$\mathbf{t}(\theta^a, \tau^a) + t - \gamma$	$\mathbf{t}(\theta^a, \tau^a) + t$	$t_\emptyset$
$w(\theta^a, \tau^a, \tau^s, \kappa)$	$\omega + \beta$	ω	ω
$x(\theta^a, \tau^a, \tau^s, \kappa)$	$\mathbf{x}(\theta^a, \tau^a)$	$\mathbf{x}(\theta^a, \tau^a)$	$x_\emptyset$

Table 1: Transfer, wage and decison for given report vector $(\theta^a, \tau^a, \tau^s)$ and bonus state τ_b.

the supervisor has an interest to report the bonus state τ_κ (to get the bonus β), but the agent wants to avoid this, because of the 'penalty' γ. We have thus created a conflict of interest, and the supervisor's private information constitutes a barrier to a solution via side-contracting. In the appendix we show that there exists a unique feasible side mechanism (satisfying (IC),(IR) and (BB)): the null side mechanism. To show this we derive lower bounds on the agent's and the supervisor's expected side payment. The supervisor's lower bound is affected by the bonus and his private knowledge of κ. He can always claim the bonus state corresponds to the true state, and thus any deviation from truthful reporting implies losing on the bonus payment. We show that adding the lower bounds yields a strictly positive expected side-payment, and thus an imbalanced budget, unless the side mechanism prescribes truthful reporting in the grand mechanism. To arrive at this conclusion the supervisor's bonus has to be sufficiently large, we require $\beta > \Delta$, where $\Delta > 0$ is the largest gain for the agent from misreporting both his type and the signal. In addition, the agent's penalty γ has to be in a certain range around $m\beta$.

To sketch the failure of side-bargaining we consider simple bribes.[31] Suppose the true signal is τ (which is common knowledge between agent and supervisor) and consider a side mechanism that specifies a joint signal report $\tau' \neq \tau$ and a side payment (henceforth bribe) b from the agent to the supervisor. A bribe $b < -\beta$ is never accepted by the supervisor, irrespective of the supervisor's knowledge about the bonus state.[32] Bribes $b \in (-\beta, 0)$ are only accepted by the supervisor who knows that $\tau_\kappa = \tau'$. The agent gains at most Δ from such a bribe, but loses $\gamma + b$, because he ends up paying the penalty γ and the bribe b. This is not profitable for the agent, whenever $\gamma > \Delta + \beta$. Bribes $b \in (0, \beta)$ are accepted by all supervisors, except for the supervisor who knows $\tau_\kappa = \tau$. Again, the agent gains at most Δ. He has to pay the

[31] The analogy is a market with asymmetric information, and the bribe is the market price. Our argument yields a no–trade result: there is no equilibrium bribe, hence everyone sticks to their outside option, in our case the truthful report in the grand mechanism.

[32] Here and in the following argument we ignore bribes for which some types of the supervisor are indifferent whether to accept it. Any probabilistic acceptance decision still leads to the desired result.

bribe and the penalty now occurs with probability $\frac{1}{m-1}$. Hence, such a bribe is not profitable whenever $\gamma > \Delta(m-1)$. Finally, bribes $b > \beta$ are accepted by the supervisor irrespective of her knowledge of κ. But such a bribe exceeds the agent's maximal gain from misreporting, because $\beta > \Delta$. We have thus established that there is no 'equilibrium' bribe, and hence, simple bribe mechanisms cannot alter truthful reporting.

The payoff structure induced by the grand mechanism resembles the payoff structure in a lemons market, c.f. Akerlof (1970).[33] There is a 'lemon supervisor', who accepts all bribes that are not too negative, but colluding with this lemon hurts the agent. However, any bribe that is accepted by non-lemon supervisors will also be accepted by this lemon. If the loss from colluding with the lemon is large enough, the agent is not willing to collude at any bribe level. Finally, there is a 'cream puff supervisor', which is the one who receives a bonus for reporting the true state while the agent pays a penalty. The agent loves colluding with the cream puff, because doing so avoids the penalty. But again, any such agreement will be accepted by the lemon such that the agent ends up paying the penalty anyways.

3.3 Fully Revealing Signals

We demonstrate the importance of Proposition 1 for the case of a fully revealing signal. From Lemma 1 and Proposition 1 we infer that *any* social choice function is implementable under collusive supervision and implementation does not require net payments to the supervisor. Which social choice functions can be implemented when the mechanism designer uses only deterministic mechanisms, i.e., the type of mechanisms that are used in the literature?[34]

With a fully informative signal we can simplify notation by considering social choice functions $(\mathbf{x}, \mathbf{t})(\theta)$. A deterministic mechanism asks the agent and the supervisor to submit a report $\theta' \in \Theta$ and uses these reports to determine the allocation. As outlined earlier, when reports do not coincide the agent and the supervisor pay a large penalty which is sufficient for deterring unilateral deviations. When reports coincide, the mechanism uses $(\mathbf{x}, \mathbf{t})$ to determine the

[33] There is one notable difference. In the classical lemons market trade is efficient for every type. Market failure typically means that only the lowest quality trades in equilibrium. In our grand mechanism trade of the lemon is not efficient (the joint payoff from colluding is strictly lower than from truthful reporting). Hence, the market failure is more severe: no trade, not even for the lowest quality, is the unique equilibrium outcome.

[34] There is no gain from using stochastic mechanisms when these do not create endogenous private information. Both the agent and the supervisor are risk-neutral with respect to money, and all randomness on alternatives x is already encoded in the choice rule $\mathbf{x} : \Theta \to \mathcal{X}$.

allocation and pays the supervisor a wage $w(\theta)$.

Proposition 2. *Suppose the signal is perfectly informative and the mechanism designer is restricted to use deterministic grand mechanisms. Under collusive supervision the SCF* $(\mathbf{x}, \mathbf{t})$ *is implementable if and only if there is a mapping* $w : \Theta \to \mathbb{R}$ *such that* $(\mathbf{x}, \mathbf{t} + w)$ *is implementable in a standard setting without supervision, i.e., whenever*

$$u(\mathbf{x}(\theta), \theta) + \mathbf{t}(\theta) + w(\theta) \geq u(\mathbf{x}(\theta'), \theta) + \mathbf{t}(\theta') + w(\theta') \qquad \forall \theta, \theta' \in \Theta. \tag{3}$$

Because colluding parties have symmetric information, they choose the joint report θ' maximizing their joint payoff. Side payments allow to split the additional surplus such that each party receives at least their reservation utility, given by truthful reporting in the grand mechanism. Hence, in any state θ the coalition selects the report θ' to maximize $u(\mathbf{x}(\theta'), \theta) + \mathbf{t}(\theta') + w(\theta')$, which yield the constraints (3). Writing $\mathbf{s}(\theta) = w(\theta) + \mathbf{t}(\theta)$, these conditions are the standard incentive compatibility conditions for implementing $(\mathbf{x}, \mathbf{s})$ in a single agent environment without supervision. While under collusive supervision with our grand mechanisms any SCF $(\mathbf{x}, \mathbf{t})$ can be implemented, upon restricting to deterministic grand mechanisms the set of implementable SCF considerably shrinks: following Rochet (1987) cyclical monotonicity of the decision rule $\mathbf{x} : \Theta \to \mathcal{X}$ is a necessary and sufficient condition for implementability without supervision, and thus (following Proposition 2) also under collusive supervision with deterministic mechanisms.

While a complete characterization of implementable SCFs under collusive supervision for arbitrary signal precision and a focus on deterministic grand mechanisms is not available, a closer look at the literature reveals that the mechanism designer is severely limited by this 'focus'. For example Celik (2009) studies a special case of our general setting with $|\Theta| = 3$ and a binary signal that partitions the type space. He shows that collusion causes an additional agency cost and it is impossible to extract the supervisor's signal for free. It is possible to avoid such agency costs, by using grand mechanisms that create endogenous asymmetric information.

3.4 An example

We conclude this section with an example to illustrate our main findings. Let $\mathcal{X} = \Theta = \{0, 1\}$ and let the agent's utility function $u(x, \theta)$ satisfy $u(x, \theta) = 0$ if $x = \theta$, and $u(x, \theta) = 10$ if $x \neq \theta$.

We are interested in implementing the social choice function $(\mathbf{x}, \mathbf{t})$ with $(\mathbf{x}, \mathbf{t})(\theta) = (\theta, t_\theta)$. The designer seeks to match the decision with the state $(x = \theta)$ and pay the agent the transfer t_θ, for some values $t_0, t_1 \in \mathbb{R}$. When only the agent observes θ but not the mechanism designer, implementing $(\mathbf{x}, \mathbf{t})$ is impossible: the agent is not willing to reveal the state θ truthfully.[35] From now on assume collusive supervision with a fully informative signal.

With deterministic mechanisms the impossibility prevails. When reports coincide on θ, the mechanism implements $x = \theta$, pays the agent t_θ and the supervisor w_θ. Any such mechanism is prone to collusion. Recall that the agent and the supervisor collude under complete information, and thus choose the joint report which maximizes the sum of their payoffs. To rule out collusion in state $\theta = 0$ it has to hold that $t_0 + w_0 \geq t_1 + w_1 + 10$. By symmetry, ruling out collusion in state $\theta = 1$ requires $t_1 + w_1 \geq t_0 + w_0 + 10$. Adding the two necessary conditions yields $0 \geq 20$, a contradiction. This confirms the result in Proposition 2, because the decision rule $\mathbf{x}$ is not monotone (and therefore not cyclically monotone).[36]

If the designer knew θ the SCF $(\mathbf{x}, \mathbf{t})$ would be implementable. Hence, Proposition 1 establishes that $(\mathbf{x}, \mathbf{t})$ can be implemented under collusive supervision. The following mechanism does the job. Define $\beta := \max\{t_0, t_1\} - \min\{t_0, t_1\} + 10 + 1$. First, draw uniformly at random a transfer schedule from the set $\{\gamma_0, \gamma_1\}$, where these schedules are given in the following table.

	γ_0	γ_1
$(t, w)(\theta = 0)$	$t_0 + \beta, -\frac{\beta}{2}$	$t_0 - \beta, \frac{\beta}{2}$
$(t, w)(\theta = 1)$	$t_1 - \beta, \frac{\beta}{2}$	$t_1 + \beta, -\frac{\beta}{2}$

Second, reveal the draw to the supervisor but not to the agent. Third, ask each player to report θ. If reports coincide, payments are made according to the drawn transfer schedule γ, and the decision matching the report is implemented. If reports do not coincide the mechanism implements some default decision and each player pays a large penalty, say -100. All of this is

[35] From the revelation principle, if $(\mathbf{x}, \mathbf{t})$ is implementable it can be implemented with a direct and truthful mechanism. The truth-telling constraints are $0 + t_0 \geq 10 + t_1$, and $0 + t_1 \geq 10 + t_0$. Adding both constraints yields $0 \geq 20$, a contradiction. Formally, the decision rule $\mathbf{x}$ is not monotone and thus not implementable.

[36] The argument continues to be valid when we allow richer message spaces, random payments, arbitrary reporting strategies and collusion in equilibrium. A collusion-proofness principle applies, which states any SCF that is implementable with such a mechanism is also implementable with a mechanism that has message set = type space, and in which players report truthfully and do not collude in equilibrium. See for exampleLaffont and Martimort (1997), Faure-Grimaud et al. (2003) and Asseyer (2019).

common knowledge – in particular, the agent knows that the supervisor knows the draw, etc. To illustrate the mechanism at work, suppose the draw results in γ_0. If both report $\theta = 0$ the decision is $x = 0$, the agent's transfer is $t_0 + \beta$, and the supervisor's wage is $-\frac{\beta}{2}$.

The example once more highlights that our construction goes beyond reducing information rents. A given decision rule $\mathbf{x}$ can be implemented with fewer information rents to the agent (without giving rise to an information rent to the supervisor). But also the entire set of implementable decision rules increases.

4 Extensions

In this section we consider several extensions to our baseline model. First, we turn to issues concerning the practical implementation of our grand mechanisms: commitment by the mechanism designer and randomization. Second, we add participation decisions for both the agent and the supervisor. Third, we study ex-ante collusion. Fourth, we extend the class of utility functions for the supervisor, allowing for limited liability and risk aversion. Finally, we briefly analyze alternative assumptions on the supervisory information, such as verifiable information and hard evidence.

4.1 Commitment

The mechanism we use for proving Proposition 1 requires randomization and confidential disclosure of the respective draw. The designer does not benefit from concealing the draw from the supervisor, because she risks that agent and supervisor reach a collusive agreement. Also, the designer does not gain from leaking this information to the agent. Hence, confidential disclosure is in the designer's own interest.

The commitment to carry out the randomization can be achieved by resorting to external randomization devices or delegating this task to a third party. There is also the possibility for providing the designer with incentives to carry out the desired randomization. To see this, recall our grand mechanism from Table 1. The decision $\mathbf{x}$ is implemented with certainty and irrespective of the drawn bonus state. Agent and supervisor receive base payments t and ω, and if they report signal τ_κ the supervisor receives a bonus while the agent pays a penalty. Some signals are more likely than others, hence the expected total payment differs across bonus

states such that the designer may wish to choose the bonus state with lowest expected payment. We account for this by adapting the grand mechanism as follows. Replace the supervisor's base wage ω by a 'bonus-state'–dependent base wage ω_κ. This adjustment does not affect the supervisor's incentives for side-contracting, hence there is still no scope for collusion. Choosing ω_κ appropriately balances expected payments across bonus states.

Proposition 3. *Any social choice function* $(\mathbf{x}, \mathbf{t})$ *that is implementable under direct supervision is also implementable under collusive supervision with zero expected payment to the supervisor, and such that the sum of expected transfer and wage is the same for each bonus state.*

4.2 Voluntary Participation

Our analysis so far focuses on the implementation of a SCF, which implicitly assumes that agent and supervisor cannot withdraw from the grand mechanism. Requiring voluntary participation does not affect our results. First, note that in our grand mechanism the agent receives the same expected utility under collusive supervision as in the benchmark case of direct supervision. Consequently, the agent is willing to participate in our grand mechanism if he was willing to participate under direct supervision. Second, we have shown that the grand mechanism implementing a given SCF $(\mathbf{x}, \mathbf{t})$ under collusive supervision pays the supervisor an expected wage of zero. Provided a reasonable outside option of zero,[37] the supervisor is willing to participate in our grand mechanism.

Adding participation decisions queries our assumption on the timing of information and collusion. It is important for our results that collusion takes place only after the supervisor learned the bonus state. Hence, this information has to be learned between the decision to participate in the grand mechanism and the first encounter with the agent. To strengthen our point we show that it is possible to endow the supervisor with private information already *before* his participation decision. In practice this amounts to offering the supervisor different (randomly selected) contracts which differ in wages. The agent knows that the supervisor is offered a contract at random, but still does not know the realized contract. Given the supervisor's outside option of zero, we have the following proposition.

[37] Any strictly positive (signal-dependent) outside option for the supervisor makes the comparison of direct and collusive supervision meaningless since the signal is costless public information under the benchmark of direct supervision. It is unclear what constitutes an appropriate (non-collusive) benchmark for collusive supervision with a signal-dependent outside option for the supervisor.

Proposition 4. *Suppose all information is revealed to the supervisor before his participation decision and the supervisor's outside option is zero. For any* $(\mathbf{x}, \mathbf{t})$ *that is implementable under direct supervision and every* $\varepsilon > 0$ *there is a grand mechanism that implements* $(\mathbf{x}, \mathbf{t})$ *and pays the supervisor an expected wage below* ε*.*

To prove Proposition 4 we adapt the grand mechanism used for proving Proposition 1 as follows. Recall, the latter mechanism pays the supervisor a base wage ω and (potentially) a bonus β. To ensure zero expected wage we have $\omega < 0$. Now, if the supervisor learns the bonus state κ and realizes that the true signal τ (which is truthfully reported in equilibrium) differs from τ_κ, he is not willing to participate. We fix this by setting $\omega = 0$. To reduce expected wages we add a new state, $\kappa = 0$, for which the supervisor's wage is zero and the agent's transfer is $\mathbf{t}(\theta, \tau) + t$ for every (θ, τ). In state $\kappa = 0$ no bonuses or penalties apply. We argue that choosing a bonus state $\kappa \neq 0$ uniformly at random with (small) probability δ, and otherwise selecting $\kappa = 0$, yields a collusion-proof grand mechanism. While this construction requires similar bonuses for the supervisor (i.e., $\beta > \Delta$ is still sufficient), it requires to increase the agent's penalty without bound. Against small bribes the supervisor is willing to collude almost certainly (except for case $\tau_\kappa = \tau$) and the likelihood of a lemon in this case is small. To render collusion unprofitable for the agent we have to increase the penalty accordingly. To bound the expected wage of the supervisor by ε we let $\delta \to 0$, i.e., the bonus accrues with vanishing probability.

4.3 Ex-ante collusion

So far we have considered interim collusion, that is collusion about reporting strategies within a grand mechanism. We next argue that our main result remains valid under ex-ante collusion. Assume collusion takes place before players decide whether to participate and that a collusive agreement is both about participation decisions and reporting strategies.[38] As in Section 4.2 we have to add outside options to make participation decisions meaningful. It is crucial for our main result that collusion takes place under asymmetric information about payments. Hence, as for Proposition 4, we endow the supervisor with private information already before his participation decision. Now, note that the lemons problem, which is the basis of our collusion–

[38] Ex-ante collusion is studied by, e.g., Pavlov (2008), Che and Kim (2009) and Mookherjee et al. (2019).

proof implementation, remains essentially unchanged under ex-ante collusion. For every signal report the coalition may agree on, there is a lemon supervisor who finds this report particularly attractive. The penalty for the agent can be chosen large, such that any collusive agreement in which the lemon participates is not attractive for the agent. It is never profitable (or feasible) to persuade the lemon to report some different signal or to not participate. In all these options his wage drops to zero, hence, also under ex-ante collusion there is only a single feasible side mechanism.[39]

4.4 Beyond risk-neutrality

The literature on collusive supervision often assumes specific preferences for the supervisor which depart from risk-neutrality.[40] Any departure from risk-neutrality exacerbates the implementation problem because it makes lotteries costly and/or prohibits (large) penalties. Nevertheless, we show that, after appropriately adapting the grand mechanism, our main result remains valid even for the two most commonly studied utility specifications for the supervisor: limited liability and risk aversion.

4.4.1 Limited Liability

We first consider the case where the supervisor is subject to limited liability:

$$U_S(x, w, \theta, \tau) = \begin{cases} w, & w \geq \hat{w}, \\ -\infty, & w < \hat{w}, \end{cases} \tag{4}$$

for some $\hat{w} \leq 0$.[41] With limited liability the mechanism cannot impose arbitrarily large negative payments on the supervisor. The grand mechanism used for proving Proposition 1 uses negative payments to the supervisor only to balance the bonus payments required for steering incentives

[39] The formal proof replicates the essential steps of the proof of Proposition 4. We have to add the option to refuse participation, which yields payoff zero for agent and supervisor. Essentially, this option is similarly attractive as sending non-conforming reports. A complete proof is available upon request from the authors.

[40] One reason is the common, but false, perception that collusion causes no problem with a risk neutral supervisor, as discussed for instance by Faure-Grimaud et al. (2003). However, this idea hinges on the restrictions to i) a binary type space and ii) a single-crossing property of the agent's preferences. Dismissing either assumption yields frictions from collusion even under risk-neutrality. For the case of i) a larger type space, we refer to Celik (2009), and for the case of ii) preferences violating single crossing just compare our Propositions 1 and 2.

[41] As before, we rule out $\hat{w} > 0$. This case is not very different, but it requires adapting the benchmark of direct supervision, since it is costly to hire the supervisor in the first place.

during side contracting. Hence, implementation per se is not affected by limited liability.

Corollary 1. *Any social choice function* $(\mathbf{x}, \mathbf{t})$ *that is implementable under direct supervision is also implementable under collusive supervision when the supervisor is subject to limited liability.*

Corollary 1 is silent about the required rent payment to the supervisor. The following proposition argues that it is possible to reduce the supervisor's rent to zero even though she is protected by limited liability as long as $\hat{w} < 0$. When $\hat{w} = 0$ the supervisor obtains a strictly positive rent, but this rent can be made arbitrarily small.

Proposition 5. *Suppose the supervisor is subject to limited liability and* $(\mathbf{x}, \mathbf{t})$ *is implementable under direct supervision. If* $\hat{w} < 0$*, the SCF* $(\mathbf{x}, \mathbf{t})$ *is implementable under collusive supervision with zero expected payments to the supervisor. If* $\hat{w} = 0$*, there is for every* $\varepsilon > 0$ *a grand mechanism implementing* $(\mathbf{x}, \mathbf{t})$ *under collusive supervision with expected payment to the supervisor below* ε*.*

For the proof we use the same mechanism as for proving Proposition 4. We add state $\kappa = 0$ for which there is no bonus and no penalty. Reducing the probability of drawing a bonus state $\kappa \neq 0$ allows for increasing the base wage ω. If $\hat{w} < 0$ we achieve zero expected wage, while for $\hat{w} = 0$ we can reduce the expected wage below any $\varepsilon > 0$.

The grand mechanisms used when accounting for the supervisor's limited liability require large penalties for the agent. As long as the supervisor can bear losses, the agent's penalty stays bounded. In particular, in the grand mechanism used for proving Proposition 1 we have that $\gamma < \beta m$, i.e., the penalty cannot be too large. If the agent is subject to limited liability, but the supervisor not, collusion-proof implementation requires reverting the roles.

Remark 2. *Suppose the designer cannot penalize the agent, i.e., the agent's final transfer cannot fall below* $\mathbf{t}(\theta, \tau) + \hat{t}$ *for any info state* (θ, τ)*. Devise a grand mechanism similar to the one used in the proof of Proposition 5, but with reversed roles: the agent knows the bonus state, but the supervisor not. Also, the agent receives the bonus* β*, while the supervisor pays the penalty* γ*. With such a mechanism we get exact implementation whenever* $\hat{t} < 0$*, and implementation with vanishing additional payments to the agent whenever* $\hat{t} = 0$*.*[42] *Note that in any case, the supervisor's expected wage is zero.*

[42]The proof mirrors the steps in the proof of Proposition 5, and is available from the authors upon request.

In general, limiting payments on one side requires inflating payments on the other side. Hence, it seems impossible to simultaneously tackle limited liability on both sides. But note that our grand mechanism is not tailored to the specific environment given by the agent's utility $u(x, \theta)$ and the desired SCF $(\mathbf{x}, \mathbf{t})$. Particular social choice functions may be implementable under collusive supervision with substantially lower bonuses and penalties.

4.4.2 Risk Aversion

We next consider a risk-averse supervisor. Formally, the supervisor's utility $U_S(x, w, \theta) = v(w)$ only depends on the wage (as before), but is not linear. To cover common examples of risk-averse preference, we only assume $v(\cdot)$ is strictly increasing, strictly concave, and satisfies $v(0) = 0$. The latter assumption is a normalization to stay in line with the direct supervision benchmark.

The case of risk aversion is similar to the case of limited liability with $\hat{w} = 0$. We use a grand mechanism that pays the supervisor a bonus with vanishing probability. As before, the bonus payments and the supervisor's private information are sufficient to break collusion. By making the probability of bonuses arbitrarily small we overcome the supervisor's risk aversion at arbitrarily small costs in the form of a small ex-ante expected wage.

Proposition 6. *Assume* $(\mathbf{x}, \mathbf{t})$ *is implementable under direct supervision, the supervisor is risk-averse and participates whenever her expected utility exceeds zero. For any* $\varepsilon > 0$*, there is a grand mechanism that implements* $(\mathbf{x}, \mathbf{t})$ *with expected payment to the supervisor below* ε*.*

4.5 Supervisory Information

Concerning supervisory information we have made two key assumptions: (i) supervisory information arrives before participating in the grand mechanism, and (ii) supervisory information is soft, i.e., parties can claim any signal irrespective of the truth. These assumptions give collusion its best chance. The literature on collusive supervision also uses different assumptions which we briefly discuss here.

Timing of information. In many applications the signal is observed only after entering a contract/mechanism. For instance, auditors are first hired (endowed with a contract) and then sent to gather information. Formally, this amounts to receiving the signal only after deciding whether to participate in the grand mechanism. The pure implementation problem is the same

as the one studied in this paper. When it comes to participation, the later arrival of information actually facilitates collusion-proof implementation as the supervisor's expected wage can be set to zero conditional on every bonus state.

Partially verifiable information. Our model assumes the signal is soft information: only cheap talk announcements regarding the signal are possible. In many realistic scenarios there is some verifiability, for instance an auditor can present detailed accounts and verifiable documents. As in Green and Laffont (1986), we can model (partial) verifiability by assuming that possible messages after signal realization τ are a subset $E(\tau) \subseteq \mathcal{T}$.[43] Effectively there are fewer deviations to consider, and thus implementation becomes simpler to achieve.

Hard information. Many studies on collusive supervision assume evidence is hard. Formally, the signal either conveys the agent's true type θ or there is no evidence at all, i.e., the signal is $\emptyset$.[44] Parties can reveal the evidence or claim not having received any, but it is impossible to fabricate false evidence. We can capture such an evidence structure as follows: let $\mathcal{T} = \{\tau_1, \ldots, \tau_n, \emptyset\}$ and assume $\pi_{ij} = 0$ for all $i \neq j$, as well as $\pi_{ii} \in [0, 1]$ and thus $\pi_{i\emptyset} = 1 - \pi_{ii} \in [0, 1]$. Hence, receiving signal τ_i is evidence for type θ_i. In addition, the message set after having received signal τ_i is given by $E(\tau_i) = \{\tau_i, \emptyset\}$, while $E(\emptyset) = \{\emptyset\}$, as in the previous paragraph on partially verifiable information. Consequently, hard evidence is subcase of partially verifiable information with a specific signal structure, and our results remain valid also for this case.

5 Conclusion

This paper studies collusion when a supervisor is employed to extract the information held by an agent. Mechanisms which endogenously create asymmetric information completely mitigate collusion. Any outcome that is implementable if the supervisor's information is public is also implementable under collusive supervision. Moreover, the supervisor does not receive an information rent. Our results help filling a gap in the literature on collusion–proof implementation. With correlated information and (sub-) coalitions of only two agents previous literature concluded that collusion is costly.

[43] Because non-conforming reports are penalized there is no loss in assuming both players have the same evidence set.

[44] See for instance Kofman and Lawarrée (1996), Kessler (2000), as well as Burlando and Motta (2015).

We show our results in a model of collusive supervision. This entails two key assumptions: (i) the supervisor has no private information vis-à-vis the agent, and (ii) the supervisor has no stake in the project, i.e., she only cares for her monetary payments. In a general setting of mechanism design with correlated private information and collusion these assumptions are restrictive. Relaxing them is thus a logical next step. Intuitively, a solution requires the grand mechanism to send private messages to each agent (not only the supervisor as in our model). The collusive side mechanism now asks each player for two pieces of information: the *exogenous* and the *endogenous* information. Proving that there is a unique feasible side mechanism becomes harder due to the complexity of the message spaces.

References

Akerlof, G. A. (1970), 'The market for "lemons": Quality uncertainty and the market mechanism', *The Quarterly Journal of Economics* **84**(3), 488.

Asseyer, A. (2019), Collusion and delegation under information control, Working paper.

Baliga, S. (1999), 'Monitoring and collusion with 'soft' information', *Journal of Law, Economics, and Organization* **15**(2), 434–440.

Börgers, T. and Norman, P. (2009), 'A note on budget balance under interim participation constraints: the case of independent types', *Economic Theory* **39**(3), 477–489.

Burlando, A. and Motta, A. (2015), 'Collusion and the organization of the firm', *American Economic Journal: Microeconomics* **7**(3), 54–84.

Celik, G. (2009), 'Mechanism design with collusive supervision', *Journal of Economic Theory* **144**(1), 69 – 95.

Che, Y.-K. and Kim, J. (2006), 'Robustly collusion-proof implementation', *Econometrica* **74**(4), 1063–1107.

Che, Y.-K. and Kim, J. (2009), 'Optimal collusion-proof auctions', *Journal of Economic Theory* **144**(2), 565–603.

Chen, J. and Micali, S. (2012), 'Collusive dominant-strategy truthfulness', *Journal of Economic Theory* **147**(3), 1300–1312.

Crémer, J. (1996), 'Manipulations by coalitions under asymmetric information: The case of Groves mechanisms', *Games and Economic Behavior* **13**(1), 39 – 73.

Faure-Grimaud, A., Laffont, J.-J. and Martimort, D. (2003), 'Collusion, delegation and supervision with soft information', *The Review of Economic Studies* **70**(2), 253–279.

Green, J. and Laffont, J.-J. (1979), 'On coalition incentive compatibility', *The Review of Economic Studies* **46**(2), 243–254.

Green, J. R. and Laffont, J.-J. (1986), 'Partially verifiable information and mechanism design', *The Review of Economic Studies* **53**(3), 447–456.

IMF (2016), Corruption: Costs and mitigating strategies, Working Paper 16/05, IMF Staff Discussion Notes.

Jeon, D.-S. and Menicucci, D. (2005), 'Optimal second-degree price discrimination and arbitrage: on the role of asymmetric information among buyers', *RAND Journal of Economics* **36**(2), 337–361.

Kessler, A. S. (2000), 'On monitoring and collusion in hierarchies', *Journal of Economic Theory* **91**(2), 280–291.

Kofman, F. and Lawarrée, J. (1993), 'Collusion in hierarchical agency', *Econometrica* **61**(3), 629–656.

Kofman, F. and Lawarrée, J. (1996), 'A prisoner's dilemma model of collusion deterrence', *Journal of Public Economics* **59**(1), 117 – 136.

Laffont, J.-J. and Martimort, D. (1997), 'Collusion under asymmetric information', *Econometrica* **65**(4), 875–912.

Laffont, J.-J. and Martimort, D. (2000), 'Mechanism design with collusion and correlation', *Econometrica* **68**(2), 309–342.

Mezzetti, C. (2004), 'Mechanism design with interdependent valuations: Efficiency', *Econometrica* **72**(5), 1617–1626.

Mookherjee, D., Motta, A. and Tsumagari, M. (2019), Consulting collusive experts, Working paper.

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

Pavlov, G. (2008), 'Auction design in the presence of collusion', *Theoretical Economics* **3**(3), 383–429.

Rahman, D. (2012), 'But who will monitor the monitor?', *American Economic Review* **102**(6), 2767–97.

Rahman, D. and Obara, I. (2010), 'Mediated partnerships', *Econometrica* **78**(1), 285–308.

Rochet, J.-C. (1987), 'A necessary and sufficient condition for rationalizability in a quasi-linear context', *Journal of Mathematical Economics* **16**(2), 191–200.

Strausz, R. (2012), 'Mediated contracts and mechanism design', *Journal of Economic Theory* **147**(3), 1280 – 1290.

Tirole, J. (1986), 'Hierarchies and bureaucracies: On the role of collusion in organizations', *Journal of Law, Economics and Organization* **2**(2), 181–214.

A Proofs of Section 3.

Proof of Lemma 1.

Follows directly from invoking a revelation principle. Note that the SCF may specify any values $\mathbf{x}(\theta,\tau),\mathbf{t}(\theta,\tau)$ for $\theta \notin \Theta(\tau)$ – these events do not occur and are thus irrelevant. □

Proof of Proposition 1.

Fix a SCF $(\mathbf{x}, \mathbf{t})$ to be implemented. Define

$$\Delta := \max_{\substack{1\leq i,l\leq n \\ 1\leq j,k\leq m}} u(\mathbf{x}(\theta_l, \tau_k), \theta_i) + \mathbf{t}(\theta_l, \tau_k) - u(\mathbf{x}(\theta_i, \tau_j), \theta_i) - \mathbf{t}(\theta_i, \tau_j) \tag{5}$$

the maximal gain for the agent from misrepresenting both the type and the signal. Consider the grand mechanism Γ with $|\Upsilon_A| = 1, \Upsilon_S = \{1, \ldots, m\}$ and $\mu(\kappa) = \frac{1}{m}$ for all $\kappa \in \Upsilon_S$. Further, set $\Sigma_A = \Theta \times \mathcal{T}, \Sigma_S = \mathcal{T}$, and

$$g(\kappa, \theta^a, \tau^a, \tau^s,) = \begin{cases} \Big(\mathbf{x}(\theta^a, \tau^a),\ \mathbf{t}(\theta, \tau^a) + t - \gamma,\ \omega + \beta\Big), & \tau^a = \tau^s = \tau_\kappa, \\ \Big(\mathbf{x}(\theta^a, \tau^a),\ \mathbf{t}(\theta, \tau^a) + t,\ \omega\Big) & \tau^a = \tau^s \neq \tau_\kappa \\ \Big(x_\emptyset,\ t_\emptyset,\ \omega\Big), & \text{else}, \end{cases} \tag{6}$$

where $\beta > \Delta,\ \omega = -\beta/m,\ t = \gamma/m$, and $\beta m - (m-1)(\beta - \Delta) < \gamma < \beta m$. Furthermore, $x_\emptyset \in \mathcal{X}$ and $t_\emptyset < \max_{\theta,\tau} u(\mathbf{x}(\theta, \tau), \theta) + \mathbf{t}(\theta, \tau) - u(x_\emptyset, \theta) - \Delta$.

We show that (a) the described grand mechanisms exhibits a non-cooperative equilibrium where both agent and supervisor report truthfully, (b) the latter equilibrium implements SCF $(\mathbf{x}, \mathbf{t})$ with expected wage zero to the supervisor, and (c) collusion does not break this equilibrium.

Step (a): Truthful reporting of the signal follows from the agent's penalty for non-conforming reports on the signal. Given that the signal is reported truthfully, the agent's expected utility from reporting θ' if his true type is θ for a given τ is

$$u(\mathbf{x}(\theta', \tau), \theta) + \mathbf{t}(\theta', \tau) + t - \tfrac{\gamma}{m} = u(\mathbf{x}(\theta', \tau), \theta) + \mathbf{t}(\theta', \tau).$$

Hence, truthful reporting of θ is optimal, because $(\mathbf{x}, \mathbf{t})$ is implementable under direct supervision.

Step (b): Truthful reporting implies that in state (θ, τ) the grand mechanism always selects $\mathbf{x}(\theta, \tau)$ and by the above arguments the agent's expected transfer is $\mathbf{t}(\theta, \tau)$. The supervisor's expected wage is $\omega + \frac{\beta}{m} = 0$.

Step (c): It remains to be shown that collusion does not break the truthful equilibrium. To show this, fix a signal realization $\tau_d \in \mathcal{T}$. The signal τ_d is common knowledge between the agent and the supervisor and thus side mechanisms condition on τ_d. To ease notation, describe a

direct side mechanism as $\Gamma = (\mathbf{p}, \mathbf{b}^a, \mathbf{b}^s)$. It consists of a set of probability vectors $\mathbf{p} = (p_{ij}^{lk})$ as well as bribe vectors $\mathbf{b}^a = (b_{ij}^a)$ and $\mathbf{b}^s = (b_{ij}^s)$. Here, p_{ij}^{lk} denotes the probability that the side mechanism demands type report θ_l and (conforming) signal report τ_k when the agent reports θ_i and the supervisor reports bonus state j. With probability $p_{ij}^{\emptyset}$ the side mechanism demands non-conforming reports on the signal (by the definition of the mechanism it is irrelevant which exact signal reports are sent in this case). Furthermore, the agent receives the bribe b_{ij}^a and the supervisor receives the bribe b_{ij}^s. Before we proceed, let us introduce some short-hand notation. For $1 \leq j, k \leq m$ denote $\mathbf{p}_j^k = \sum_{i=1}^n \sum_{l=1}^n \pi_i^d p_{ij}^{lk}$ the probability that the side mechanism triggers conforming signal report τ_k when the supervisor reports bonus state j and the agent reports his type truthfully. Recall $\pi_i^d = Prob(\theta_i|\tau_d)$ is the conditional probability of type θ_i when the signal is τ_d. Similarly, denote $\mathbf{p}_j^{\emptyset} = \sum_{i=1}^n \pi_i^d p_{ij}^{\emptyset}$ the probability that the side-mechanism triggers a non-conforming signal-report after report j by the supervisor and truthful report by the agent. Further, define $b_i^a = \sum_{j=1}^m \frac{1}{m} b_{ij}^a$, resp. $b_j^s = \sum_{i=1}^n \pi_i^d b_{ij}^s$ the expected bribe to the agent of type i, resp. the supervisor of type j. Denote $B^a = \sum_i \pi_i^d b_i^a$ and $B^s = \sum_j \frac{1}{m} b_j^s$ the ex-ante expected bribe payments. We continue with a Lemma characterizing incentive compatible side mechanisms.

Lemma A.1. *In any incentive compatible side mechanism*

$$\mathbf{p}_j^j + \mathbf{p}_k^k \geq \mathbf{p}_k^j + \mathbf{p}_j^k, \qquad \forall 1 \leq j, k \leq m. \tag{7}$$

Proof. From the supervisor's IC-constraints we have that $b_j^s + \omega + \mathbf{p}_j^j \beta \geq b_k^s + \omega + \mathbf{p}_k^j \beta$ for all j, k. Summing constraints for $j \mapsto k$ and $k \mapsto j$, and using $\beta > 0$ yields the desired inequalities. □

We next establish a lower bound on bribe payments received by the supervisor.

Lemma A.2. *In any incentive compatible and individually rational side mechanism*

$$B^s \geq (1 - \mathbf{p}_d^d)\beta + \frac{\beta}{m} \sum_{j \neq d} \left(\mathbf{p}_d^j - \mathbf{p}_j^j \right). \tag{8}$$

Proof. The supervisor who knows $\kappa = d$ participates in the side mechanism whenever $b_d^s + \omega + \mathbf{p}_d^d \beta \geq \omega + \beta$, which yields $b_d^s \geq (1 - \mathbf{p}_d^d)\beta$. The supervisor who knows $\kappa = j$ prefers a truthful report over reporting $\kappa' = d$, whenever $b_j^s + \omega + \mathbf{p}_j^j \beta \geq b_d^s + \omega + \mathbf{p}_d^j \beta$. After rearranging, we get

$b_j^s \geq b_d^s + \left(\mathbf{p}_d^j - \mathbf{p}_j^j\right)\beta$. Hence,

$$B^s = \frac{1}{m}\sum_{j=1}^{m} b_j^s \geq b_d^s + \frac{\beta}{m}\sum_{j\neq d}\left(\mathbf{p}_d^j - \mathbf{p}_j^j\right) \geq (1-\mathbf{p}_d^d)\beta + \frac{\beta}{m}\sum_{j\neq d}\left(\mathbf{p}_d^j - \mathbf{p}_j^j\right).$$

□

We next establish a lower bound on the bribe payments to the agent.

Lemma A.3. *In any incentive compatible and individually rational side mechanism*

$$B^a \geq \frac{\gamma}{m}\left(\sum_{j=1}^{m}\mathbf{p}_j^j - 1\right) - \frac{\Delta}{m}\sum_{j=1}^{m}\sum_{k\neq d}\mathbf{p}_j^k + \frac{\Delta}{m}\sum_{j=1}^{m}\mathbf{p}_j^\emptyset. \tag{9}$$

Proof. The agent of type θ_i participates whenever

$$\begin{aligned} b_i^a + \sum_{j=1}^{m}\frac{1}{m}p_{ij}^\emptyset\Big(u(x_\emptyset,\theta_i)+t_\emptyset\Big) + \sum_{j=1}^{m}\sum_{l=1}^{n}\sum_{k=1}^{m}\frac{1}{m}p_{ij}^{lk}\Big[u(\mathbf{x}(\theta_l,\tau_k),\theta_i)+\mathbf{t}(\theta_l,\tau_k)\Big] + t - \sum_{j=1}^{m}\sum_{l=1}^{n}p_{ij}^{lj}\frac{\gamma}{m} \\ \geq u(\mathbf{x}(\theta_i,\tau_d),\theta_i) + \mathbf{t}(\theta_i,\tau_d) + t - \frac{\gamma}{m}. \end{aligned}$$

Hence,

$$\begin{aligned} B^a &= \sum_{i=1}^{n}\pi_i^d b_i^a \\ &\geq \frac{\gamma}{m}\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{n}\pi_i^d p_{ij}^{lj} - 1\right\} + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\pi_i^d}{m}p_{ij}^\emptyset\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d) - u(x_\emptyset,\theta_i) - t_\emptyset\Big] \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{n}\sum_{k=1}^{m}\frac{\pi_i^d}{m}p_{ij}^{lk}\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d) - u(\mathbf{x}(\theta_l,\tau_k),\theta_i) - \mathbf{t}(\theta_l,\tau_k)\Big] \\ &= \frac{\gamma}{m}\left\{\sum_{j=1}^{m}\mathbf{p}_j^j - 1\right\} + \frac{1}{m}\sum_{j=1}^{m}\mathbf{p}_j^\emptyset\underbrace{\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d) - u(x_\emptyset,\theta_i) - t_\emptyset\Big]}_{\geq\,\Delta} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{n}\frac{\pi_i^d}{m}p_{ij}^{ld}\underbrace{\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d) - u(\mathbf{x}(\theta_l,\tau_d),\theta_i) - \mathbf{t}(\theta_l,\tau_d)\Big]}_{\geq\,0} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{n}\sum_{k\neq d}\frac{\pi_i^d}{m}p_{ij}^{lk}\underbrace{\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d) - u(\mathbf{x}(\theta_l,\tau_k),\theta_i) - \mathbf{t}(\theta_l,\tau_k)\Big]}_{\geq\,-\Delta} \\ &\geq \frac{\gamma}{m}\left\{\sum_{j=1}^{m}\mathbf{p}_j^j - 1\right\} + \frac{\Delta}{m}\sum_{j=1}^{m}\mathbf{p}_j^\emptyset - \frac{\Delta}{m}\sum_{j=1}^{m}\sum_{k\neq d}\mathbf{p}_j^k. \quad \square \end{aligned}$$

Combining (8) and (9), a side mechanism that satisfies (IC) and (IR) requires total expected

budget B of at least

$$\begin{aligned} B &\geq (1-\mathbf{p}_d^d)\beta + \frac{\beta}{m}\sum_{j\neq d}\left(\mathbf{p}_d^j - \mathbf{p}_j^j\right) + \frac{\gamma}{m}\left(\sum_{j=1}^m \mathbf{p}_j^j - 1\right) - \frac{\Delta}{m}\sum_{j=1}^m\sum_{k\neq d}\mathbf{p}_j^k + \frac{\Delta}{m}\sum_{j=1}^m \mathbf{p}_j^\emptyset \\ &= \left(\beta - \tfrac{\gamma}{m}\right)\left(1-\mathbf{p}_d^d\right) + \tfrac{\Delta}{m}\sum_{\substack{j<k\\ j,k\neq d}}\left[\mathbf{p}_j^j + \mathbf{p}_k^k - \mathbf{p}_j^k - \mathbf{p}_k^j\right] + \tfrac{\gamma-\beta-(m-1)\Delta}{m}\sum_{j\neq d}\mathbf{p}_j^j + \tfrac{\beta-\Delta}{m}\sum_{j\neq d}\mathbf{p}_d^j + \tfrac{\Delta}{m}\sum_{j=1}^m \mathbf{p}_j^\emptyset. \end{aligned}$$

Recall, that by assumption $\beta > \Delta > 0$ and $\gamma \in (\beta m - (m-1)(\beta-\Delta), \beta m)$. Hence, to guarantee an ex-ante balanced budget the side mechanism has to satisfy (i) $\mathbf{p}_d^d = 1$, (ii) $\mathbf{p}_j^j + \mathbf{p}_k^k = \mathbf{p}_j^k + \mathbf{p}_k^j$ for all $j, k \neq d$ (recall Lemma A.1), (iii) $\mathbf{p}_j^j = 0$ for all $j \neq d$, (iv) $\mathbf{p}_d^j = 0$ for all $j \neq d$, and (v) $\mathbf{p}_j^\emptyset = 0$ for all j. Properties (ii),(ii) and (v) imply that $\mathbf{p}_j^d = 1$ for all $j \neq d$. Together with (i) this implies there is a *unique* side mechanisms that satisfies (IC),(IR) and (BB): the null side mechanism prescribing a truthful report of the signal τ_d without exchanging side payments.

In steps (a) – (c) we have thus shown that the described mechanisms implements $(\mathbf{x}, \mathbf{t})$ under collusive supervision, while paying the supervisor an expected wage of zero. □

Proof of Proposition 2.

Follows from applying standard results from the literature. Note that under collusive supervison with deterministic mechanisms we can invoke a collusion–proofness principle, see for instance (Faure-Grimaud et al., 2003, p. 273). Their proof for a supervisor with a CARA utility function and an agent whose preferences are linear in his type and the alternative directly carries over to our setting if only deterministic grand mechanisms are feasible.

B Proofs of Section 4.

Proof of Proposition 3.

Devise a mechanism similar to the one used for proving Proposition 1. For each bonus state $\kappa = 1, \ldots, m$ define the same payments, except for replacing ω in supervisor's wage by ω_κ. Hence, the supervisor's base wage directly depends on the bonus state. Note that the incentive structure remains unaffected. Consequently, Lemma A.1 applies and we obtain the same lower bound on the supervisor's expected side payment B^s. Because the agent's transfer and the

allocation are unchanged, the lower bound on the agent's side payment B^a is not affected either. Hence, there is a unique feasible side mechanism, and this side mechanism prescribes truthful reporting in the grand mechanisms without any side payments.

In this new grand mechanism, for any bonus state κ the expected total monetary payment is

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\Big(\mathbf{t}(\theta_i,\tau_j)+t+\omega_\kappa\Big)+\sum_{i=1}^{n}\pi_{i\kappa}\Big(\beta-\gamma\Big)=\mathbb{E}(\mathbf{t}(\theta,\tau))+t+\omega_\kappa+(\beta-\gamma)\sum_{i=1}^{n}\pi_{i\kappa}.$$

Setting $\omega_\kappa=(\gamma-\beta)\sum_{i=1}^{n}\pi_{i\kappa}-t$ yields the desired result. Note that this mechanism still pays the supervisor an expected wage of zero:

$$\sum_{\kappa=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\pi_{ij}}{m}\left(\omega_\kappa+\mathbb{I}_{\{j=\kappa\}}\beta\right)=\frac{\gamma-\beta}{m}\sum_{\kappa=1}^{m}\sum_{i=1}^{n}\pi_{i\kappa}-t+\frac{\beta}{m}\sum_{\kappa=1}^{m}\sum_{i=1}^{n}\pi_{i\kappa}=\frac{\gamma}{m}-t=0.$$

An intermediate result for proving Propositions 4 to 6.

We define a class of mechanisms that allows us to prove Propositions 4 to 6. Fix $(\mathbf{x},\mathbf{t})$ and $\delta\in(0,1)$. Let $\Delta>0$ be defined as in the proof of Proposition 1. Define a grand mechanism with $\Upsilon_S=\{0,1,\ldots,m\}$, $|\Upsilon_A|=1$ and $\mu(0)=1-\delta$, as well as $\mu(\kappa)=\frac{\delta}{m}$ for $\kappa\neq 0$. Further, set $\Sigma_A=\Theta\times\mathcal{T},\Sigma_S=\mathcal{T}$, and

$$g(,\kappa,\theta^a,\tau^a,\tau^s)=\begin{cases}\Big(\mathbf{x}(\theta^a,\tau^a),\ \mathbf{t}(\theta,\tau^a)+t-\gamma,\ \omega+\beta\Big), & \tau^a=\tau^s=\tau_\kappa,\\ \Big(\mathbf{x}(\theta^a,\tau^a),\ \mathbf{t}(\theta,\tau^a)+t,\ \omega\Big) & \tau^a=\tau^s\neq\tau_\kappa\\ \Big(x_\emptyset,\ t_\emptyset,\ \omega\Big), & \text{else,}\end{cases}\tag{10}$$

where $\beta>\Delta$, $t=(\delta\gamma)/m$, and $\frac{\beta m}{\delta}-\frac{m-\delta}{\delta}(\beta-\Delta)<\gamma<\frac{\beta m}{\delta}$. The values $t_\emptyset$ and $x_\emptyset$ are chosen as in the proof of Proposition 1.

Proposition B.1. *The grand mechanism described above implements* $(\mathbf{x},\mathbf{t})$ *under collusive supervision.*

Proof. The proof repeats the steps for proving Proposition 1. It is straightforward to verify that if $(\mathbf{x},\mathbf{t})$ is implementable under direct supervision the described mechanism has an equilibrium in which both agent and supervisor report their information truthfully. Also, this equilibrium implements $(\mathbf{x},\mathbf{t})$. It remains to be shown that collusion does not break this non-cooperative

equilibrium. We follow the strategy used for proving Proposition 1 in showing that there is no feasible side mechanism other than non-cooperative play.

Fix a true signal $\tau_d \in \mathcal{T}$. We first derive some useful inequalities.

Lemma B.1. *In any incentive compatible side mechanism*

$$\mathbf{p}_j^j + \mathbf{p}_k^k \geq \mathbf{p}_j^k + \mathbf{p}_k^j, \qquad \forall j,k \in \{1,\dots,m\}, \tag{11}$$

and

$$\mathbf{p}_j^j \geq \mathbf{p}_j^0, \qquad \forall j \in \{1,\dots,m\}. \tag{12}$$

Proof. From the supervisor's IC we get $\mathbf{p}_j^j\beta + \mathbf{b}_j^s \geq \mathbf{p}_k^j\beta + \mathbf{b}_k^s$ for all $j,k \in \{1,\dots,m\}$. Pairwise comparison of these inequalities for $j,k \neq 0$ yields (11). The supervisor of type 0 prefers reporting truthfully whenever $\mathbf{b}_0^s \geq \mathbf{b}_j^s$ for all $1 \leq j \leq m$. Adding the incentive constraints of type j and 0 yields (12). □

Next, we derive lower bounds on the expected side payments to the agent and the supervisor.

Lemma B.2. *In any incentive compatible and individually rational side mechanism*

$$B^s \geq \sum_{j=1}^{m} \frac{\delta}{m}\mathbf{b}_j^s + (1-\delta)\mathbf{b}_0^s \geq \left(1 - \mathbf{p}_d^d\right)\beta + \frac{\delta}{m}\beta\sum_{j\neq d}\left(\mathbf{p}_d^j - \mathbf{p}_j^j\right). \tag{13}$$

Proof. The supervisor of type d participates in the side mechanism whenever $\mathbf{p}_d^d\beta + \mathbf{b}_d^s \geq \beta$, or equivalently whenever $\mathbf{b}_d^s \geq (1-\mathbf{p}_d^d)\beta$. From the respective incentive constraints we have $\mathbf{b}_j^s \geq \mathbf{b}_d^s + (\mathbf{p}_d^j - \mathbf{p}_j^j)\beta$, and $\mathbf{b}_0^s \geq \mathbf{b}_d^s$. Summing the derived lower bounds on b_j^s yields (13). □

Next consider the agent. Using only participation constraints, we get the following lemma.

Lemma B.3. *In any incentive compatible and individually rational side mechanism*

$$B^a \geq -\frac{\delta}{m}\Delta\sum_{j=1}^{m}\sum_{k\neq d}\mathbf{p}_j^k - (1-\delta)\Delta\sum_{j\neq d}\mathbf{p}_0^j + \Delta\left(\frac{\delta}{m}\sum_{j=1}^{m}\mathbf{p}_j^{\emptyset} + (1-\delta)\mathbf{p}_0^{\emptyset}\right) + \frac{\delta}{m}\gamma\left(\sum_{j=1}^{m}\mathbf{p}_j^j - 1\right). \tag{14}$$

Proof. Type θ_i participates in the side-mechanism, whenever

$$\begin{aligned}
&\mathbf{b}_i^a + \sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{m}\frac{\delta}{m}p_{ij}^{kl}\Big[u(\mathbf{x}(\theta_k,\tau_l),\theta_i)+\mathbf{t}(\theta_k,\tau_l)\Big] + (1-\delta)\sum_{k=1}^{n}\sum_{l=1}^{m}p_{i0}^{kl}\Big[u(\mathbf{x}(\theta_k,\tau_l),\theta_i)+\mathbf{t}(\theta_k,\tau_l)\Big]\\
&+\left\{\frac{\delta}{m}\sum_{j=1}^{m}p_{ij}^{\emptyset}+(1-\delta)p_{i0}^{\emptyset}\right\}\Big[u(x_\emptyset,\theta_i)+t_\emptyset\Big]+t-\sum_{j=1}^{m}\sum_{k=1}^{n}p_{ij}^{kj}\frac{\delta}{m}\gamma\\
&\geq u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)+t-\frac{\delta}{m}\gamma
\end{aligned}$$

Rearranging and summing over i yields

$$\begin{aligned}
B^a &= \sum_{i=1}^{n}\pi_i^d b_i^a\\
&\geq \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{m}\pi_i^d\tfrac{\delta}{m}p_{ij}^{kl}\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)-u(\mathbf{x}(\theta_k,\tau_l),\theta_i)-\mathbf{t}(\theta_k,\tau_l)\Big]\\
&\quad+(1-\delta)\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{m}\pi_i^d p_{i0}^{kl}\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)-u(\mathbf{x}(\theta_k,\tau_l),\theta_i)-\mathbf{t}(\theta_k,\tau_l)\Big]\\
&\quad+\left\{\frac{\delta}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\pi_i^d p_{ij}^{\emptyset}+(1-\delta)\sum_{i=1}^{n}\pi_i^d p_{i0}^{\emptyset}\right\}\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)-u(x_\emptyset,\theta_i)-t_\emptyset\Big]\\
&\quad+\frac{\delta}{m}\gamma\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{n}\pi_i^d p_{ij}^{kj}-1\right)\\
&\geq -\frac{\delta}{m}\Delta\sum_{j=1}^{m}\sum_{l\neq d}\mathbf{p}_j^l-(1-\delta)\Delta\sum_{j\neq d}\mathbf{p}_0^j+\Delta\left(\frac{\delta}{m}\sum_{j=1}^{m}\mathbf{p}_j^{\emptyset}+(1-\delta)\mathbf{p}_0^{\emptyset}\right)+\frac{\delta}{m}\gamma\left(\sum_{j=1}^{m}\mathbf{p}_j^j-1\right).
\end{aligned}$$

□

Using (13) and (14), a side mechanism that satisfies (IC) and (IR) requires total budget

$$\begin{aligned}
B^s+B^a &\geq \left(1-\mathbf{p}_d^d\right)\beta+\frac{\delta}{m}\beta\sum_{j\neq d}\left(\mathbf{p}_d^j-\mathbf{p}_j^j\right)\\
&\quad-\frac{\delta}{m}\Delta\sum_{j=1}^{m}\sum_{k\neq d}\mathbf{p}_j^k-(1-\delta)\Delta\sum_{j\neq d}\mathbf{p}_0^j+\Delta\left(\frac{\delta}{m}\sum_{j=1}^{m}\mathbf{p}_j^{\emptyset}+(1-\delta)\mathbf{p}_0^{\emptyset}\right)+\frac{\delta}{m}\gamma\left(\sum_{j=1}^{m}\mathbf{p}_j^j-1\right)\\
&= (1-\mathbf{p}_d^d)\left(\beta-\gamma\tfrac{\delta}{m}\right)+\frac{\delta}{m}(\beta-\Delta)\sum_{j\neq d}\mathbf{p}_d^j+\Delta\left(\frac{\delta}{m}\sum_{j=1}^{m}\mathbf{p}_j^{\emptyset}+(1-\delta)\mathbf{p}_0^{\emptyset}\right)\\
&\quad+\left(\tfrac{\delta}{m}\gamma-\tfrac{\delta}{m}\beta-\tfrac{(m-1)\delta}{m}\Delta-(1-\delta)\Delta\right)\sum_{j\neq d}\mathbf{p}_j^j+\tfrac{\delta}{m}\Delta\sum_{\substack{0<j<k\\ j,k\neq d}}\left(\mathbf{p}_j^j+\mathbf{p}_k^k-\mathbf{p}_j^k-\mathbf{p}_k^j\right)\\
&\quad+(1-\delta)\Delta\sum_{j\neq d}\left(\mathbf{p}_j^j-\mathbf{p}_0^j\right)
\end{aligned}$$

As in the proof of Proposition 1 all terms above are non-negative and have to be zero to satisfy

(BB). Thus, $\mathbf{p}_d^d = 1$; $\mathbf{p}_j^\emptyset = 0$ for all $j = 0, 1, \ldots, m$; $\mathbf{p}_j^j = \mathbf{p}_0^j = 0$ for all $j \neq d$. Consequently, $\mathbf{p}_0^d = 1$. Furthermore, (11) implies $\mathbf{p}_j^d = 1$ for all $j \neq d$. We have thus shown there is a unique side mechanism that satisfies (IC),(IR) and (BB): the null side mechanism. Hence, for every $\delta \in (0, 1)$ our grand mechanism implements $(\mathbf{x}, \mathbf{t})$ under collusive supervision. □

Proof of Proposition 4.

Consider the class of grand mechanisms from Proposition B.1. Set $\omega = 0$ and fix δ. Then all wages to the supervisor are non-negative. Consequently, the supervisor participates even after observing the realization of the bonus state. The supervisor's expected wage payment is $\sum_{\kappa=1}^{m} \sum_{i=1}^{n} \frac{\delta}{m} \pi_{i\kappa} \beta = \frac{\delta}{m} \beta$. Recall, β is finite, hence chosing $\delta < \frac{m\varepsilon}{\beta}$ yields expected wage to the supervisor below ε. □

Proof of Proposition 5.

For $\hat{w} = 0$ we can use the same mechanism that we used for proving Proposition 4. When $\hat{w} < 0$, set $\omega = \hat{w}$ and choose δ such that

$$0 = (1 - \delta)\hat{w} + \sum_{\kappa=1}^{m} \sum_{i=1}^{n} \frac{\delta}{m} \pi_{i\kappa} \beta = (1 - \delta)\hat{w} + \tfrac{\delta}{m} \beta,$$

which yields $\delta = \frac{-m\hat{w}}{\beta - m\hat{w}} \in (0, 1)$. □

Proof of Proposition 6.

The proof uses the same construction as for proving Proposition 4. During the proof we have to account for the supervisor's risk preferences, but all steps are essentially identical. Having $\delta \to 0$ the riskiness in the supervisor's payment vanishes. We omit the details. □