Discussion Paper Series – CRC TR 224

Discussion Paper No. 019
Project B 01

Sweet Lemons: Mitigating Collusion in Organizations

Colin von Negenborn [1]
Martin Pollrich [2]

May 2018

[1] Humboldt University Berlin,Department of Economics, Spandauer Str. 1, 10178 Berlin, Germany;
e-mail: von.negenborn@hu-berlin.de
[2] University of Bonn, Department of Economics, Lennéstr. 37, 53112 Bonn, Germany;
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]

May 23, 2018

Abstract

This paper shows that the possibility of collusion between an agent and a supervisor imposes no restrictions on the set of implementable social choice functions (SCF) and associated payoff vectors. Any SCF and any payoff profile that are implementable if the supervisor's information was public is also implementable when this information is private and collusion is possible. To implement a given SCF we propose a one-sided mechanism that endogenously creates private information for the supervisor vis-à-vis the agent, and conditions both players' payoffs on this endogenous information. We show that in such a mechanism all collusive side-bargaining fails, similar to the trade failure in Akerlof's (1970) car market and in models of bilateral trade.

Keywords: Mechanism Design, Collusion, Asymmetric Information, Correlation

JEL Codes: D82, D83, L51

[a]Humboldt University Berlin, Department of Economics, Spandauer Str. 1, 10178 Berlin, Germany; e-mail: von.negenborn@hu-berlin.de

[b]University of Bonn, Department of Economics, Lennéstr. 37, 53112 Bonn, Germany; 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 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 funding by the DFG through CRC TR 224, the Hausdorff Center of Mathematics in Bonn, and the DAAD.

1 Introduction

Ever since Akerlof (1970) economists are concerned with asymmetric information as a significant friction in economic interactions, severely impeding the implementation of efficient outcomes. In response, institutions are established whose major role lies in reducing informational asymmetries. Examples range from auditors verifying financial statements of companies to certifiers asserting the (hidden) quality of products. The value of these institutions rests on their credibility to truthfully reveal the information they observe. In this paper we are concerned with collusion as one potential threat to credibility and as a hindrance to implementation.[1] A privately informed party may seek to bribe the intermediary to forge reports, e.g. paying an auditor to conceal unfavorable evidence or a certifier for releasing favorable product information. Is it possible to design institutions that give intermediaries incentives to reveal all their information truthfully, while at the same time making them resistant to the threat of collusion?

Addressing the issue of collusion in the framework of mechanism design is not straightforward. A comprehensive analysis of the (detrimental) effects of collusion requires considering *all conceivable* mechanisms, i.e. all institutional frameworks. In standard mechanism design the revelation principle facilitates this task, by allowing to focus on incentive compatible direct mechanisms. However, collusion introduces cooperative aspects and thus precludes invoking a revelation principle. The previous literature tacitly invokes a (quasi) revelation principle: a mechanism merely asks for reports on the initially observed information. These restrictions for instance ignore additional information such as the knowledge about collusive agreements.[2]

In this paper we use a more general class of mechanisms that not only asks the players for reports on their information but also provides them with discriminatory information. The latter feature endogenously creates (additional) asymmetric information between the possibly colluding parties. We show that an appropriate design of the mechanism, in particular designing the information structure and monetary payments, allows to completely overcome

[1] The threat of collusion 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.

[2] The few exceptions include Chen and Micali (2012), who allow for richer message spaces: agents are also asked which coalition they belong to. Similar to our approach, Ortner and Chassang (2018) use mechanisms that create endogenous asymmetric information, but in a setting with moral hazard and incomplete contracts. See the literature review below for a more detailed account.

collusion: any outcome that is implementable were the third party information credible can also be implemented when the third party is prone to colluding.

Our model builds on and extends existing models of collusive supervision, as initiated by Tirole (1986). There is a privately informed agent and a supervisor who receives a (possibly noisy) signal on the agent's information. The agent derives utility from the implemented alternative and monetary transfers, while the supervisor only cares about money. We are interested in implementing social choice functions (SCF), mapping states of the world into outcomes. The benchmark is given by the scenario where the supervisor's signal is public. Our main result shows, first, that *any* SCF which is implementable under such direct supervision is also implementable under collusive supervision, i.e. if the designer cannot observe the signal and covert side agreements between the agent and the supervisor are possible. Second, implementation does *not* require net payments to the supervisor and hence does not bring additional costs vis-à-vis the benchmark. Neither agent nor supervisor receive an extra information rent stemming from their possibility to collude and the signal hence is extracted for free.

We prove this result by constructing a mechanism which endogenously creates asymmetric information between agent and supervisor, thus driving a wedge between the collusive parties. Collusion, after all, is a bilateral bargaining problem regarding the reports to be sent to the grand mechanism. It has been noted in the literature that asymmetric information makes bargaining more difficult, up to a complete failure of reaching an agreement. While this effect is detrimental in the used car market of Akerlof (1970), it is good news in the case of collusive side bargaining. Our mechanism provides the supervisor with payoff relevant information she holds *privately* vis-à-vis the agent. This is achieved by randomizing the transfers to agent as well as supervisor and, in addition, sending a private signal to the supervisor on the realization of these transfers. At the collusion stage, this induces a bargaining situation with interdependent valuations. The agent does not know whether he faces a 'lemon' supervisor, who is easy to bribe but unattractive to collude with – or whether he faces a 'cream puff', where collusion is highly attractive since a forged report strongly increases payments. The grand mechanism we construct this way renders side bargaining basically impossible: any side mechanism which specifies reports other than truthtelling can not at the same time be incentive compatible, individually rational and ex-ante budget balanced. We thus induce the desired non-cooperative Bayesian Nash equilibrium as the *unique feasible* outcome of

any collusive agreement.

The practical implementation of our mechanisms requires two aspects: randomization and differential information. First, regarding randomization, the mechanisms we propose do not differ from stochastic mechanisms as analyzed in contract theory and mechanism design. Many real-life processes of supervision are subject to randomness, such as auditors only checking random samples of the managerial data. Second, we introduce endogenous asymmetric information. Most audit scenarios provide scope for such a differential distribution of knowledge: a manager may know the auditor's general wage structure but not her specific wage class, he may be unaware whether the auditor was randomly sent for a general investigation or whether certain data triggered the search for specific information, and in a firm with several managerial divisions he may not even realize whether he himself is subject to the audit or one of the other managers. Each of these scenarios precludes the manager from exact knowledge of an optimal bribing level and hence impedes collusion.

The results we present are robust to various changes. We can incorporate standard assumptions on the supervisor's utility function, in particular risk aversion and limited liability as well as voluntary participation. If these constraints become severe – e.g. by precluding any ex post extractions from the supervisor –, we can still *virtually* implement the direct supervision benchmark: by providing incentive payments in some of the contracts offered to the players and having the probabilities of these contracts go to zero, the mechanism designer can obtain the desired information with rent payments vanishing in the limit. The same robustness holds for changes to the timing, where the supervisor decides on participation in the grand mechanism *after* having obtained her private signal which is used to create an asymmetry between herself and the agent.

Our results have further implications for the theory of collusive supervision: first, we add to the debate on delegated vs. centralized contracting in a hierarchy prone to collusion. The mechanism we propose is necessarily centralized and delegation is strictly dominated. When using only direct mechanisms, Faure-Grimaud et al. (2003) demonstrate that the optimal mechanism can be implemented via delegation to the supervisor. We also contribute to the question as to whether such a restriction to direct mechanism is in fact appropriate in the presence of collusion: as argued above, the literature has – with only few exceptions – tacitly invoked a revelation principle by only considering mechanisms where agent and supervisor are asked to report their information. As a result, collusion was thought to be costly in that

it limits the set of implementable social choice functions and gives rise to information rents to the supervisor. Our results imply that these findings are the result of an undue restriction on the class of permitted mechanisms, not a consequence of collusion per se. Finally, we add to the analysis on optimal supervisory information. Faure-Grimaud et al. (2003) as well as Asseyer (2018) argue that an imperfectly informed supervisor is preferable: the value of supervisory information increases in informativeness, but the collusive rents are higher the better informed the supervisor is. However, we show that using appropriate non-direct mechanisms the supervisory signal can be extracted for free. Hence, a more informative signal is always beneficial.

Further implications regard collusive settings where there exists ex-ante two-sided asymmetric information, such as in auctions: in our setting, there initially only is one-sided asymmetry in that the agent has an informational advantage over the supervisor. In an auction setting, however, there may be several agents, each possessing private information e.g. on the value of the object to be auctioned. It is an avenue for further research to analyze how our results extend to such mechanism design problems.

The paper proceeds as follows: after reviewing the literature in Section 2, we present a toy example in Section 3, illustrating the notion of non-direct mechanisms we propose. The full model is introduced in Section 4. In Section 5 we present our main results, including a comparison of collusive supervision to the benchmark of direct supervision. Section 6 extends the baseline model into various dimensions, such as risk aversion, limited liability or participation constraints. Section 7 concludes. All proofs are in the Appendix.

2 Related Literature

This paper contributes to three strands of literature on the theory of collusion. First, we add to the analysis of collusion in vertical hierarchies, comprised of a principal–supervisor–agent framework. Second, we shed further light on the use of mechanism design in the context of collusion and, in particular, on the (non-)applicability of the revelation principle under cooperative play. Third, we exploit the mechanism designer's possibility to endogenously create asymmetric information and thus add to previous work doing so. We discuss each strand in more detail.

Tirole (1986) introduces the three-tier principal–supervisor–agent model as the workhorse model for studying collusion in organizations. In such a hierarchy the agent has some private information, and the supervisor has more information on the agent than the principal. The focus lies on the transmission of information from the supervisor to the principal, when collusion between agent and supervisor is possible. Tirole shows that supervision is beneficial despite collusion, but collusion gives rise to additional distortions (both productive and rents).

The subsequent literature studies variations and extensions of Tirole's model. Kofman and Lawarrée (1993), Kofman and Lawarrée (1996*a*), Kessler (2004) and Burlando and Motta (2015) keep Tirole's assumption of hard evidence, i.e. evidence can only be concealed and not forged. Another strand studies collusion with soft supervisory information, where parties can claim any realized signal, e.g. Baliga (1999), Faure-Grimaud et al. (2003), Celik (2009), and Asseyer (2018). In almost all contributions collusion is shown to be detrimental. Only Kessler (2004) and Burlando and Motta (2015) provide instances where collusion can be overcome, but their results crucially depend on the fact that supervisors learn their information – and thus send their reports – only after the agent communicated.[3] Throughout, this literature focuses on mechanisms where the agent and the supervisor send reports only. In contrast, we allow for a richer class of mechanisms and show that collusion can be prevented at no cost irrespective of the timing of information, preferences of the players and the properties of supervisory information (hard vs soft).

The problem of collusion has also attracted much attention for general mechanism design problems. Already Green and Laffont (1979) point out that Vickrey–Clarke–Groves mechanisms are vulnerable against coalition formation.[4] Laffont and Martimort (1997, 2000) provide analyses of specific mechanism design problems with independent and correlated information, concluding implementability is limited in the presence of collusion. Che and Kim (2006) address the problem of collusion in a general mechanism design setting, showing that collusion is not an issue when the agents' information is independent. For a given mechanism that yields revenue R absent collusion, they construct a new mechanism by re-designing payments such that revenue is constant at R, irrespective of whether and which coalition between the agents forms. Their results continue to hold for correlated

[3]Their results also depend on the agent's preferences which have to satisfy a single crossing property.
[4]See also Crémer (1996) and Chen and Micali (2012).

information—as is necessarily the case in our model of supervision—but only when there are at least three agents and only with respect to the grand coalition.

We complement the analysis of Che and Kim (2006) by showing there exist mechanisms that prevent collusion even in models with correlated information and only two players. While Che and Kim embrace collusion via making the mechanism designer indifferent as to whether collusion takes place, our construction aims at making it impossible for the agent and the supervisor to form a coalition in the first place. Both approaches use asymmetric information as a friction in side bargaining. As Che and Kim (2006) show, with independent information these frictions prevent any coalition from reaching outcomes that cannot be obtained in the absence of collusion. With correlated information the players have an informational advantage over the designer, requiring the addition of informational asymmetries to break collusion. Doing so requires using mechanisms that go beyond those usually employed in mechanism design – with Chen and Micali (2012) being an exception, as argued above –, where informed parties only report their information.

Finally, there is a small literature using mechanisms that endogenously create asymmetric information between agents. Rahman and Obara (2010) exploit this in the team problem with moral hazard and budget-balanced payments. They introduce a mediator who sends confidential recommendations to the group members, and condition transfers directly on these recommendations. As in our mechanism, after receiving their recommendations agents are asymmetrically informed as to the consequences of their actions. Strausz (2012) relates these *mediated contracts* back to the revelation principle. Rahman (2012) studies the problem of providing a monitor with incentives to actually monitor. A mediated mechanism, secretly recommending the agent to sometimes shirk, provides the monitor with incentives. Rahman's mechanism is prone to collusion: agent and monitor may side contract to report false evidence. Combining his construction with ours yields a mechanism that both provides the monitor with incentives to monitor and at the same prevents collusion.

In a setting more closely related to ours, Ortner and Chassang (2018) prove the benefit of creating endogenous asymmetric information when fighting collusion. They randomize the supervisor's wage and provide only the supervisor – not the agent – with information on the realization. The colluding parties therefore have different expectations on the value of collusion. As a result, the agent does not know the bribe required to corrupt the supervisor and an informational friction impedes their side bargaining. However, we depart from this

model in several ways. Most importantly, the authors use a setting of moral hazard and incomplete contracts: while the detection of collusion only affects the supervisor's wage (and not the agent's), the agent's choice is only relevant for his own payoff (and not the supervisor's). That is, transfers specified by the contract do not allow for a full mechanism design approach in that they are not contingent on all of the reports or observables. As a result, the authors cannot exploit the full interdependence of valuations that we use in the present paper, where the private information sent to the supervisor is payoff-relevant for *both* of the collusive parties. In their model, collusion therefore remains costly, though they can reduce these costs by the use of randomized wages. The randomized mechanism we propose, however, *fully* mitigates the costs of collusion by randomizing not only over wages, but over whole mechanisms instead.

3 Illustrative Example

Consider a company that wants to build a new branch but requires a manager to do so. There are two possible locations, x_1 and x_2, and two possible states of the world, θ_1 and θ_2. The company would like to implement a specific social choice function (SCF) (x, t) which pins down a location x_i and a transfer t_i to the manager for each state θ_i. In particular, it seeks to build the branch at the location matching the state without paying a rent to the supervisor: $x_i = \theta_i$ and $t \equiv 0$. The manager has idiosyncratic preferences given by the following table and an outside option of zero:

u	θ_1	θ_2
x_1	0	10
x_2	10	0

If the state θ is publicly known, this SCF can readily be implemented since the manager receives his outside option and has no private information. The same holds true if the company cannot observe the state directly but hires a supervisor who honestly reports it: since the supervisor – as opposed to the manager – does not gain utility from the chosen location, she reports the state truthfully. Matters become more intricate if the manager can bribe the supervisor to make a false report. Is the implementation of the SCF impeded by the possibility of collusion?

Most of the literature focuses on what we later define as *deterministic contracts*: both

the manager and the supervisor are merely asked for a report on the sate of the world, which pins down a chosen location x, a transfer t to the manager and a wage w to the supervisor. In equilibrium they report truthfully[5] and we can thus denote by (t_i, w_i) the payments given a joint report θ_i on the state. Now if the true state is θ_1, the manager may offer a bribe b to the supervisor such that they jointly report θ_2 instead. The agent will offer a bribe of at most $b \leq t_2 - t_1 + 10$, while the supervisor will require at least $b \geq w_1 - w_2$. A necessary condition for the contract to induce truthelling hence is that no bribe exists which meets both of these requirements, i.e. we need $t_2 - t_1 + 10 < w_1 - w_2$. The same logic applies in state θ_2. Adding the respective truthelling constraint to the one just derived yields $20 < 0$, a contradiction. Hence, no simple contract exists which implements the SCF $(\mathbf{x}, \mathbf{t})$ introduced above.

We now propose a more general mechanism which achieves the task of implementing $(\mathbf{x}, \mathbf{t})$. We again denote by (t_i, w_i) the payments given a joint report θ_i. Consider a menu of simple contracts given by $\Gamma = \{\gamma_0, \gamma_1, \gamma_2\}$, where

	γ_0	γ_1	γ_2
t_1, w_1	$22, -11$	$-22, 11$	$0, 0$
t_2, w_2	$22, -11$	$0, 0$	$-22, 11$

Assume each contract is equally likely and penalties are large for non-unanimous reports on θ. Crucially, the supervisor is informed which of the three contracts is realized (after she has accepted to participate), while the manager is not. When choosing their reports and considering collusion, they therefore face asymmetric information.

It is readily verified that neither manager nor supervisor have an incentive to unilaterally deviate, and that in expectation both receive their outside option of zero when reporting truthfully. In particular, $\mathrm{E}[t|\theta_i] = 0$ is satisfied for all i. Is there scope for collusion, i.e. a coordinated misreport? Suppose we are in state θ_1, and the manager offers a bribe b to the supervisor for jointly reporting state θ_2. A bribe $b < -11$ is never accepted by the supervisor, and thus not a threat. Bribes $b \in (-11, 0)$ are only accepted by the supervisor of 'type' γ_2, i.e. the supervisor who knows contract γ_2 was selected. But the manager's utility reduces from zero to $\frac{1}{3}(0+22)+\frac{1}{3}(0-22)+\frac{1}{3}(10-22-b) = -\frac{1}{3}(12+b) < 0$. Next consider bribes $b \in (0, 11)$. Both the γ_0- and the γ_2-supervisor find such bribes acceptable. The manager's expected payoff from offering such a bribe is $\frac{1}{3}(10+22-b)+\frac{1}{3}(0-22)+\frac{1}{3}(10-22-b) = -\frac{2}{3}(1+b) < 0$.

[5] A version of the collusion-proofness principle applies here, see Faure-Grimaud et al. (2003).

Lastly, consider bribes $b > 11$, which all supervisor-types accept. The expected profit is $\frac{1}{3}(10+22)+\frac{1}{3}(10+0)+\frac{1}{3}(10-22)-b = 10-b < 0$.

Hence, there is no bribe that manager and supervisor can agree on to coordinate their misreport. As we show in this paper, this impossibility extends to more general forms of collusion: there is nor side mechanism which is ex-ante budget balanced, acceptable for both manager and supervisor and different from truthful reporting. Using more general mechanisms than mere deterministic contracts allows the firm to fully overcome collusion.

4 Model

The basic setting. There is a single agent with type $\theta \in \Theta = \{\theta_1, \ldots, \theta_n\}$. The agent's preferences over alternatives $x \in \mathcal{X}$ and monetary transfers $t \in \mathbb{R}$ are given by

$$U_A(x,t,\theta) = u(x,\theta) + t. \tag{1}$$

We shall make no restrictions on the function $u(\cdot,\cdot)$, other than it being real-valued. The set $\mathcal{X}$ of alternatives is arbitrary as well.[6] Additive separability and risk-neutrality with respect to monetary payments are crucial assumptions as will become clear from our analysis. They are, however, standard in the literature on collusion theory.

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 in that it does not (directly) enter the agent's utility function. However, it may contain information about the agent's type. We allow for arbitrary correlation between θ and τ. Formally, the prior is given by

$$\pi_{ij} = \Pr(\theta = \theta_i, \tau = \tau_j), \quad 1 \le i \le n, 1 \le j \le m. \tag{2}$$

We assume full support, i.e. for all $1 \le i \le n$ there is a j such that $\pi_{ij} > 0$, and vice versa.[7] From the unconditional distribution we deduce the conditional distribution of the agent's type given the signal. Let π_i^j denote the probability that the agent's type is θ_i conditional

[6] In particular we require neither compactness nor convexity.

[7] The full support assumptions rules out redundant types and signals, and is thus a mere matter of convenience.

on the realized signal $\tau = \tau_j$. From Bayes' rule we have

$$\pi_i^j = \Pr(\theta = \theta_i | \tau = \tau_j) = \frac{\pi_{ij}}{\pi_{1j} + \ldots + \pi_{nj}}. \tag{3}$$

We are interested in implementing social choice functions

$$(\mathbf{x}, \mathbf{t}) \ : \ \Theta \times \mathcal{T} \to \mathcal{X} \times \mathbb{R}, \tag{4}$$

mapping the agent's type θ and the signal τ into a decision $\mathbf{x}(\theta, \tau) \in \mathcal{X}$ and a monetary 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 of all, we do not rule out payoff-relevance for outsiders, as this does not conflict with the implementation problem itself. Second, even if an 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).

Information and Supervision. Throughout we assume only the agent observes his type θ. Regarding the signal τ we distinguish two scenarios.

- *Direct supervision:* the signal τ is public.
- *Collusive supervision:* both a third party – the supervisor – and the agent privately observe τ. In addition, agent and supervisor can collude, as specified below. The supervisor's utility only depends on the monetary wage $w \in \mathbb{R}$ she receives,

$$U_S(x, w, \theta) = w. \tag{5}$$

  Hence, the supervisor has no intrinsic motive to misreport information, in particular the signal τ is also irrelevant for the supervisor's payoff.[8] As the supervisor does not observe θ, we face a nested information structure: the agent's type θ is observed only by the agent, but neither the supervisor nor any outsider. The signal τ is observed both by the agent and the supervisor, but not by anyone else. Hence, the agent has private information vis-à-vis the supervisor, and both the supervisor and the agent have private information vis-à-vis any outsider or the mechanism designer.

[8] Hence, collusion is the only reason for why the supervisor misreports the signal she observes. We discuss the issue of an interested supervisor – whose payoff may depend on x, θ and τ – in Section 6.4.1.

Mechanisms. In order to implement an SCF $(\mathbf{x}, \mathbf{t})$, we can construct arbitrary mechanisms, each corresponding to an extensive form game. Depending on the scenario described above, this game is played by the agent alone (in the scenario of direct supervision) or by both agent and supervisor (in the scenario of collusive supervision). In the latter case, the possibility of cooperation prevents us from invoking a revelation principle.[9] We are thus left with the universe of mechanisms to choose from. Nevertheless, to prove our main results it will turn out that it is without loss to confine attention to a limited class of mechanisms.

We distinguish between *grand mechanisms* on the one hand, used to implement an SCF as explained above, and *side mechanisms* on the other hand, governing collusion as introduced below. In the following we describe two classes of grand mechanisms: *deterministic contracts* and *one-sided randomized menus*.

Definition 1. *A* deterministic contract $\gamma = (x, t, w)$ *consists of*

(i) a mapping $x : \Theta \times \mathcal{T} \times \mathcal{T} \to \mathcal{X}$*, which maps every report vector into an alternative,*

(ii) a mapping $t : \Theta \times \mathcal{T} \times \mathcal{T} \to \mathbb{R}$*, which maps every report vector into a payment to (or from, if negative) the agent, and*

(iii) a mapping $w : \Theta \times \mathcal{T} \times \mathcal{T} \to \mathbb{R}$*, which maps every report vector into a payment to (or from, if negative) the supervisor.*

A *deterministic contract* γ asks the agent to report a type and a signal, and the supervisor to report a signal. For each report vector, γ specifies an alternative and payments. In the absence of collusion there would be no loss of generality in restricting the mechanism designer to the use of deterministic mechanisms because both the agent and the supervisor are risk-neutral with respect to monetary payments, and randomization over alternatives x is incorporated by having $\mathcal{X}$ also contain lotteries over deterministic alternatives.

We emphasize the following notation: bold characters denote a social choice function, see $\mathbf{x}$ and $\mathbf{t}$ in eq. (4), mapping the state space $\Theta \times \mathcal{T}$ into an alternative and a payment to the agent, respectively. Non-bold characters x and t denote the elements of a deterministic contract, where the domain is given by the set of report profiles $\Theta \times \mathcal{T} \times \mathcal{T}$ instead, see Definition 1. When we analyze the implementability of a social choice function $(\mathbf{x}, \mathbf{t})$ below,

[9] As discussed by Laffont and Martimort (2000), a revelation principle in a setting with collusion would have to include the possibility for asking the agents which side mechanism they are playing. On top of that, the mechanism would give recommendations which side mechanism to form. Consequently, invoking a revelation principle does not give rise to a tractable framework.

we effectively ask if there is a contract $\gamma = (x, t, w)$ that supports it.[10]

The second class of grand mechanisms is defined as follows.

Definition 2. *A* one-sided randomized menu (Γ, q) *consists of*

(i) a (finite) set $\Gamma = \{\gamma_1, \ldots, \gamma_k\}$ *of deterministic contracts, and*

(ii) a probability distribution $q \in \Delta(\Gamma)$*, where* $q(\gamma_i) > 0$ *for all* $1 \leq i \leq k$,[11]

with the following timing:

(1) a deterministic contract γ' *is selected from* Γ *according to* q*,*

(2) the supervisor gets privately informed of the selected deterministic contract γ'*, and*

(3) the agent reports a type and a signal, and the supervisor reports a signal. The report vector determines the outcome according to γ'*.*

One important assumption we make is that the supervisor receives no information that is verifiable vis-à-vis the agent. For example, a supervisor's contract may entail several wage classes, but the actual wage is not written into the contract, though known to the supervisor. All details of a one-sided randomized menu are common knowledge: the agent knows that the deterministic contracts are drawn according to q from the menu Γ, and that the supervisor knows the realization; the supervisor knows that the agent knows and so on. Mechanisms of this kind, which are random and reveal only parts to its agents, have been previously employed in problems of moral hazard, for instance by Rahman and Obara (2010) and Rahman (2012).

Collusion. The vast majority of work on mechanism design assumes that agents act non-cooperatively when 'playing' the mechanism. We want to challenge this assumption by allowing the agent and the supervisor to coordinate their reports to the grand mechanism. Formally, we model collusion taking place when the agent and the supervisor decide about their reporting strategies. Collusion describes the formation of a side contract that commits players to a particular reporting strategy and the exchange of monetary payments. Given

[10] We do not include the wage w paid to the supervisor in the notation of a SCF $(\mathbf{x}, \mathbf{t})$ since we are interested in comparing the implementability of such SCFs under 'direct supervision' (where the signal τ is public and hence neither a supervisor nor a wage w is required) to the scenario of collusive supervision, where the wage is used as an auxiliary mean to implement $(\mathbf{x}, \mathbf{t})$. We will nevertheless address the question whether this implementation causes additional costs, i.e. rent payment $\mathrm{E}[w] > 0$ to the supervisor.

[11] The assumption $p(\gamma_i) > 0$ for all i is clearly without loss. It avoids cases where the menu contains redundant items.

the above definition of grand mechanisms, there are two sources of private information at the collusion stage. First, the agent knows his type, thus has (residual) private information vis-à-vis the supervisor. Second, in a one-sided randomized menu, the supervisor in turn also has private information vis-à-vis the agent: she knows the selected simple contract, and thus the exact payoffs following each joint report. The signal realization τ, on the other hand, is common knowledge between the agent and the supervisor, though unknown to any outsider and thus to the grand mechanism.

We want to allow arbitrary formations of a side mechanism, thus modeling collusion in the most general way possible and to the largest hindrance of implementing social choice functions. We therefore follow Laffont and Martimort (1997) in assuming that a disinterested third party proposes a collusive side mechanism Γ^c.[12] Without loss of generality we can focus on direct side mechanisms in which all parties report their private information truthfully. Formally, as side mechanism is described by $\Gamma^c = \big(\Theta, \Gamma, o\big)$: the agent reports a type $\theta' \in \Theta$, the supervisor reports a simple contract $\gamma' \in \Gamma$, and the outcome function $o : \Theta \times \Gamma \to \Delta(\Theta \times \mathcal{T} \times \mathcal{T}) \times \mathbb{R} \times \mathbb{R}$ determines the (potentially random) joint report to the grand mechanism $(\theta^a, \tau^a, \tau^s)$ and monetary side-payments to the agent and the supervisor. We assume players have no access to outside sources of money, and therefore restrict side-payments to be ex-ante balanced-budget, i.e. the expected sum of side payments is non-positive. Hence, we allow for burning money and ex-post imbalances in the side mechanism's budget. Lastly, no player can be forced to enter a side mechanism. Note that the outside option is endogenous: it is given by the (expected) payoff from playing the grand mechanism non-cooperatively. At this point we shall also assume *passive* beliefs after unilateral rejection of the side mechanism, i.e. players do not update their beliefs about the other after an unanticipated rejection of the side mechanism. To summarize, collusive side mechanisms are direct mechanisms that are incentive compatible, ex-ante budget balanced, and satisfy individual rationality, i.e. voluntary participation.

Timing. When a one-sided randomized menu is used and collusion is possible, events unfold as follows (if a deterministic contract is used instead, stage 2 below is obsolete, hence the timing with a deterministic contract is the special case consisting only of stages 1, 3–5):

1. Nature draws (θ, τ) and informs the agent about θ, and both the supervisor and the

[12] Note this in particular includes take-it-or-leave-it offer from either of the collusive parties.

agent about τ.

2. Nature draws γ' from Γ according to q, and confidentially reveals γ' to the supervisor.

3. 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 5, otherwise to stage 4.

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

5. Both the agent and the supervisor send their reports to the grand mechanism (in case of an effective side-contract the reports determined therein), and the allocation is determined according to γ'.

5 Analysis

In this section we present our main result on the implementability of social choice functions under collusive supervision. Two benchmark cases – direct supervision and non-collusive supervision – provide the foundation for our main result, which is presented in Proposition 1. At the end of this section we use the special case of a fully informative signal to provide a particular illustration of our main results: we compare the set of implementable SCFs when imposing a restriction to deterministic contracts to the case when no such restriction is made.

5.1 Direct Supervision

Let us first consider the case of direct supervision, where the signal τ is publicly observable. 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.

Given that τ is public, a mechanism can condition on the realized signal. Furthermore, we can invoke a revelation principle and focus without loss of generality on direct mechanisms in which the agent truthfully reports his type. A social choice function $(\mathbf{x}, \mathbf{t})$ can be directly associated with a direct mechanism, which asks the agent to report a type θ' and determines an allocation using the reported type and the publicly observable signal τ according to

$(\mathbf{x}, \mathbf{t})$. Note that a direct mechanism allows the agent to report only types that have strictly positive probability under the realized signal, i.e. the report set after signal realization τ is $\Theta(\tau) \coloneqq \{\theta \in \Theta | \Pr(\theta, \tau) > 0\}$. To ensure truthful reporting by the agent, the usual incentive constraints arise, as stated in the following Lemma:

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

$$u(\mathbf{x}(\theta, \tau), \theta) + \mathbf{t}(\theta, \tau) \geq u(\mathbf{x}(\theta', \tau), \theta) + \mathbf{t}(\theta', \tau) \quad \forall \theta, \theta' \in \Theta(\tau) \ \forall \tau \in \mathcal{T}. \tag{6}$$

Note the special case of a fully informative signal. Formally, $\mathcal{T} = \{\tau_1, \ldots, \tau_n\}$ 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 (6) has no bite. Consequently *any* social choice function $(\mathbf{x}, \mathbf{t})$ is implementable.[13] We will return to fully informative signals at the end of this section.

5.2 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 – that is, players behave non-cooperatively.

Fix an SCF $(\mathbf{x}, \mathbf{t})$ that is implementable under direct supervision. Implementation under non-collusive supervision exploits the fact that both the agent and the supervisor are symmetrically informed about the signal, and that the agent has no incentives to misreport his type provided the signal is reported truthfully. A classic way of implementing a given SCF $(\mathbf{x}, \mathbf{t})$ (that is implementable under direct supervision) uses *shoot-the-liar* mechanisms.[14] Such a mechanism is a deterministic contract γ, asking the agent to report his type and the signal, and the supervisor to report the signal. If the reports on the signal coincide, the allocation is chosen according to $(\mathbf{x}, \mathbf{t})$ using the reported signal and the agent's reported type. Otherwise, if reported signals are non-unanimous, some alternative $x \in \mathcal{X}$ is chosen together with large negative monetary payments. The latter 'penalties' deter unilateral deviations from truthfully reporting the signal, and implementability under direct supervision implies

[13] Yet, in this case participation constraints, which we do not explicitly model in this section, do constrain feasible allocations. For an analysis including these constraints, see section Section 6.1.

[14] For a complete characterization of implementable SCFs in two-agent environments see Moore and Repullo (1990).

the agent has no incentive to misreport his type. Note that such a mechanism does not require any payment to the supervisor *in equilibrium*. In fact, setting the supervisor's wage equal to zero for all conforming reports on the signal, and to some arbitrarily large negative value otherwise, yields the desired implementation. The following Lemma summarizes.

Lemma 2. *Any social choice function* $(\mathbf{x}, \mathbf{t})$ *that is implementable under direct supervision is also implementable under non-collusive supervision without paying the supervisor, i.e. there is a deterministic contract* $\gamma = (x, t, w)$ *such that*

$$x(\theta, \tau, \tau) = \mathbf{x}(\theta, \tau),\ t(\theta, \tau, \tau) = \mathbf{t}(\theta, \tau), \quad \textit{and} \quad w(\theta, \tau, \tau) = 0, \tag{7}$$

and truthful reporting is an equilibrium of γ.

5.3 Collusive Supervision

We now move on to study the general setting with collusive supervision. While shoot-the-liar mechanisms prove effective in deterring *unilateral* deviations, they can in general not deter *group* deviations. Since τ is symmetrically known by the agent and the supervisor, *any* unilateral deviation can be identified easily,[15] allowing for effective punishment. Moreover, since all unilateral deviations occur off the equilibrium path, these punishments do not conflict with on-path implementation of the social choice function.

However, deterrence becomes more complicated with collusion. A collusive side mechanisms gives players the possibility to coordinate their reports, allowing for joint deviations. This makes it impossible to unambiguously identify all potential deviations, and using only monetary penalties that arise off-path is not sufficient for deterring all joint deviations. In general, 'direct' mechanisms struggle with this task, as has been pointed out for instance by Che and Kim (2006), who show that a direct mechanism lacks instruments to simultaneously deter all unilateral and all group deviations.[16] But when it comes to joint deviations, effective deterrence does not necessarily require rendering all such deviations unattractive for all group members *simultaneously*. Since the group has to agree on a deviation strategy,

[15] Strictly speaking, it can be identified *whether* there was a unilateral deviation, but the *deviant's idendity* cannot be identified.

[16] Their approach embraces group deviations by 'selling' the firm to the agents. However, they state that with only two agents (as in our model) "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) The solution they propose therefore is not applicable in our setting.

it is sufficient to drive a wedge between the colluding parties that makes it impossible to reach such an agreement in the first place. That is, agent and supervisor should be caused to disagree *when* it is profitable to collude, and *how* to do so. In the following we use mechanisms that exactly achieve this goal, by making it impossible to agree on any collusive sided-mechanism that differs from non-cooperative equilibrium play.

To this end we will use one-sided randomized menus (Γ, q), as introduced in Definition 2. Recall that these consist of a set $\Gamma = (\gamma_1, \ldots, \gamma_k)$ of deterministic contracts $\gamma_i = (x_i, t_i, w_i)$ and a probability distribution $q \in \Delta(\Gamma)$, where $q(\gamma)$ denotes the probability that deterministic contract $\gamma \in \Gamma$ is selected. However, their key property is one-sidedness: the supervisor knows which deterministic contract was selected, while the agent does not. We can thus use information design to deter joint deviations.

We are now in the position to state our main result:

Proposition 1. *Any social choice function* $(\mathbf{x}, \mathbf{t})$ *that is implementable under direct supervision is also implementable under collusive supervision without net payments to the supervisor. Formally, for any SCF* $(\mathbf{x}, \mathbf{t})$ *that is implementable under direct supervision, there is a one-sided randomized menu* (Γ, q)*, such that*

i) $x_i(\theta, \tau, \tau) = \mathbf{x}(\theta, \tau)$ *for all* $(\theta, \tau) \in \Theta \times \mathcal{T}$ *and all* i

ii) $\sum_i q(\gamma_i) t_i(\theta, \tau, \tau) = \mathbf{t}(\theta, \tau)$ *for all* $(\theta, \tau) \in \Theta \times \mathcal{T}$,

iii) $\sum_i q(\gamma_i) w_i(\theta, \tau, \tau) = 0$ *for all* $(\theta, \tau) \in \Theta \times \mathcal{T}$,

and that exhibits an equilibrium in which all reports are truthful, despite collusion.

Contrary to conventional wisdom about implementation in the presence of collusion, Proposition 1 shows that collusion neither restricts implementability nor does it give rise to additional (coalitional) information rents. In other words, despite collusion it is possible to extract the signal τ for free and implement the SCF $(\mathbf{x}, \mathbf{t})$ as if τ was publicly known. Our finding challenges previously obtained insights on optimal mechanisms in principal–supervisor–agent hierarchies under the threat of collusion, for instance in Faure-Grimaud et al. (2003), Celik (2009) and Asseyer (2018). In particular, we show that it is always better to have a well-informed supervisor. Previous work argued that there is an interior optimum to the precision of the signal τ, since a perfectly informative signal caused supervisor and agent to have symmetric information, supposedly making it hard to break up their coalition. Using our mechanism, the signal can be extracted for free and hence is more beneficial

the more precise it is. Before providing an intuition for the proof, we make an additional remark: our mechanism implements the desired selection of an alternative $\mathbf{x}(\theta, \tau)$ exactly, but the payments are implemented in expectation, i.e. the agent's ex-ante expected transfer is $\mathbf{t}(\theta, \tau)$, and the supervisor's ex-ante expected wage is zero.[17] As before, our mechanism uses penalties after non-conforming reports to deter unilateral deviations.

The proof of Proposition 1 proceeds as follows. For a given SCF $(\mathbf{x}, \mathbf{t})$ we first define $\Delta > 0$ as the largest gain the agent could receive from (mis-)reporting the information in some state (θ, τ), claiming some other state (θ', τ').[18] Using Δ, we construct a menu of deterministic contracts $\Gamma = \{\gamma_0, \gamma_1, \ldots, \gamma_m\}$ as follows. For $1 \leq i \leq m$ the contract γ_i rewards the supervisor with a bonus $\Delta + \kappa$ in case both players report signal τ_i, where κ is an arbitrary positive number.[19] At the same time, the agent is *penalized* by an amount $y\Delta$, where $y > 1$ is specified in the proof. In every other case these contracts pay the supervisor zero, and the agent according to $\mathbf{t}(\cdot, \cdot)$. The contract γ_0, on the other hand, is constructed such that all expected payments add up to zero in case of the supervisor, and to $\mathbf{t}(\theta, \tau)$ in case of the agent, if all deterministic contracts are equally likely. Finally, all deterministic contracts apply sufficiently large penalties after non-conforming reports on the signal.

Now consider the one-sided randomized menu (Γ, q), where $q(\gamma_i) = 1/(m+1)$ as mentioned before. First of all, penalties after non-conforming reports on the signal deter unilateral deviations, hence there is a non-cooperative equilibrium where both the agent and the supervisor report their information truthfully – just as in the case of non-collusive supervision. Can collusion break this non-cooperative equilibrium? Consider a collusive side-mechanism with side payments. Such a side mechanism asks the supervisor for a report on the realized deterministic contract γ, and the agent for a report on his type θ. Using the supervisor's incentive and participation constraints, we then derive a lower bound on the expected bribe required in any feasible side mechanism. Similarly, we derive a lower bound for the agent's expected bribe, using only his participation constraints. Recall that a side mechanism cannot print money, i.e. the expected sum of bribes paid out must not exceed zero. Adding up the two lower bounds, we then show that there is a *unique* side

[17] Here risk-neutrality with respect to monetary payments is crucial. Section 6.3 shows that our results are robust when allowing for arbitrary risk-attitudes of the supervisor.

[18] Whenever $\Delta \leq 0$ implementation is not an issue, as the agent himself would have an incentive to truthfully reveal all information.

[19] Strictly speaking, κ can be zero. Whenever $\kappa > 0$ there is a unique feasible side-mechanism, and this side mechanism corresponds to the non-cooperative equilibrium.

	γ_1	$\cdots$	γ_k	$\cdots$	γ_m	γ_0
w	0	$\cdots$	$\Delta + \kappa$	$\cdots$	0	$-(\Delta + \kappa)$
t	$\mathbf{t}(\theta, \tau_k)$	$\cdots$	$\mathbf{t}(\theta, \tau_k) - y\Delta$	$\cdots$	$\mathbf{t}(\theta, \tau_k)$	$\mathbf{t}(\theta, \tau_k) + y\Delta$

Table 1: Payments after joint report (θ, τ_k) in the respective deterministic contracts.

mechanism that is incentive compatible, individually rational and ex-ante balanced budget: the mechanism that specifies the non-cooperative equilibrium without side payments.

Why is there no other feasible side mechanism than this one? For every signal realization there is a contract promising the supervisor a large bonus for reporting the signal truthfully. Thus, in order to move away from truthfulness, the side mechanism needs to promise a bribe to the respective 'type' of supervisor in order to compensate her for the foregone bonus. Here, a 'type' indicates the supervisor's private, payoff-relevant information on the realized deterministic contract γ. But this positive bribe is also attractive for the supervisor when she knows the selected contract does not pay a bonus for reporting the true signal. The bribe she demands is proportional to $\Delta + \kappa$ and weighted by the probabilities which the side mechanism assigns to the different report profiles. This demand level, however, outweighs the agent's potential gain from engaging in collusion by the mere definition of Δ.[20]

Another way to see this is to consider the formation of a side mechanisms as the formation of an equilibrium bribe for reporting some τ_k, while the true signal is τ_l. Table 1 lists the monetary payments from jointly reporting signal τ_k under the respective deterministic contracts. The situation is reminiscent of Akerlof's 1970 car market: the supervisor (aka the seller) knows the exact payments after any joint report, while the agent (the buyer) does not. Large bribes, higher than $\Delta + \kappa$, attract any supervisor, but are unattractive for the agent by the definition of Δ. Intermediate bribes, $0 < b < \Delta + \kappa$, attract all supervisors but 'type' γ_l. However, it attracts type γ_k which implies a penalty of size $y\Delta$ for the agent and hence is a 'lemon' from the agent's perspective, in the sense of Akerlof. For large enough y also this bribe is thus unattractive for the agent. Similarly, negative bribes cannot be equilibrium values. Consequently, there is *no* equilibrium bribe. This holds for any deviating report, thus only truthful reports with a bribe of zero remain.[21] This intuition highlights

[20]The factor y in the agent's transfer stems from the fact that he updates the value of collusion conditional on the supervisor accepting a bribe. That is, he forms a new posterior on the realized contract knowing that the supervisor would not have accepted if the contract offering her a bonus for truthful reporting was drawn. For a detailed analysis we refer to the formal proof.

[21]The result for general mechanisms mirrors an insight by Samuelson (1984), who shows that in a lemons market there is *some* mechanism with a positive level of trade if and only if the there is a positive level of trade when the uninformed party (the buyers) make price offers.

the resemblance to the lemons problem. The lemon is the supervisor receiving a bonus for reporting a particular signal τ', differing from the true signal τ. It is impossible to separate this supervisor from the others, when intending to report τ' instead of τ. But payoffs are constructed such that the agent can never gain when collusion includes this type of supervisor, as in the car market where trade fails due to the buyer's inability to separate lemons from good cars.

5.4 Benefits of Randomized Menus

When the signal is fully informative we can exactly quantify the gain from using one-sided randomized menus, compared to a restriction to deterministic contracts. Recall that with a fully informative signal the agent and the supervisor are symmetrically informed about θ, and we can thus omit explicitly denoting τ. An SCF $(\mathbf{x}, \mathbf{t})$ thus only conditions on θ. By Lemma 1 and Proposition 1 *any* social choice function is implementable under collusive supervision.

What happens when we restrict to deterministic contracts under collusive supervision? A deterministic contract specifies for any (conforming) report on θ an alternative $x(\theta)$, a transfer to the agent $t(\theta)$ and a wage to the supervisor $w(\theta)$. Side bargaining takes place under symmetric information, i.e without frictions. Effectively, the coalition chooses the (joint) report θ^r that maximizes their joint payoff $u(x(\theta^r), \theta) + t(\theta^r) + w(\theta^r)$. Side payments are used to split the cake to ensure each party receives at least their reservation utility (given by truthful reporting in the grand mechanism). But then, the coalition acts as if it was a single entity with preferences equal to the agent's. It is well known that in this case only SCFs can be implemented whose alternative selection $\mathbf{x}(\cdot)$ is cyclically monotone.

Lemma 3. *Under collusive supervision with a perfectly informative signal, and if only deterministic contracts are used, an SCF* $(\mathbf{x}, \mathbf{t})$ *is implementable if and only if* $\mathbf{x}$ *is cyclically monotone, i.e. if and only if for every sequence of length* $k \in \mathbb{N}$ *of types* $(\theta^1, \dots, \theta^k) \in \Theta^k$ *with* $\theta^k = \theta^1$*, we have*

$$\sum_{l=1}^{k-1} u(\mathbf{x}(\theta^l), \theta^{l+1}) - u(\mathbf{x}(\theta^l), \theta^l) \leq 0. \tag{8}$$

A complete characterization of implementable SCFs under collusive supervision, i.e. for arbitrary signal precision, when restricting to deterministic contracts is not available. Celik (2009) provides an example with $|\Theta| = 3$ and a binary signal that partitions the

type space. He shows that even when only monotonic SCFs are considered (for which the alternative selection $\mathbf{x}(\cdot)$ can be implemented with some transfer rule $\mathbf{t}$), there is an additional 'coalitional information rent' involved. It stems from the possibility of coordinating the agent's and the supervisor's report and is *in addition* to individual rents due to unilateral deviation. Hence, in Celik's setting the signal cannot be obtained for free. Our Proposition 1 shows that using one-sided randomized menus allows for strictly reducing the information rents in this example.

6 Extensions

This section provides several extensions to our baseline model. First, we introduce voluntary participation and outside options for both the agent and the supervisor. Second, we discuss an alternative timing assumption for our grand mechanisms. Third, we study various alternatives for the supervisor's (and the agent's) preferences that appear in the literature, such as limited liability and risk-aversion. Finally, we briefly analyze some aspects of supervisory information (timing of information arrival, partially verifiable signals, hard evidence), and discuss the underlying commitment assumptions in our grand mechanisms.

6.1 Voluntary Participation

In the previous section our focus was on implementation only: both the agent and the supervisor were forced to participate in the grand mechanism. However, our results are robust to the introduction of voluntary participation decisions, adding participation constraints to the problem.

Regarding the agent we can allow for arbitrary type- and signal-dependent outside options $\bar{u}(\theta, \tau)$. To satisfy the agent's outside option we now have to add the usual participation constraints to the benchmark of direct supervision, implying that the agent's utility in every state (θ, τ) exceeds his outside option: $u(\mathbf{x}(\theta, \tau), \theta) + \mathbf{t}(\theta, \tau) \geq \bar{u}(\theta, \tau)$. Note that the one-sided randomized menus used to prove Proposition 1 implement $(\mathbf{x}, \mathbf{t})$ with the same expected transfer to the agent. Consequently, the agent's interim expected utility equals $u(\mathbf{x}(\theta, \tau), \theta) + \mathbf{t}(\theta, \tau)$ and thus exceeds the value of his outside option in every state (θ, τ).

A sensible outside option for the supervisor is zero. Since the supervisor's (expected) wage equals zero in every state (θ, τ), she is willing to participate in the mechanisms we

use for proving Lemma 2 and Proposition 1. In the same way, any weakly negative outside option (that may also depend on the state (θ, τ)) does not affect our results. A strictly positive outside option renders supervision costly, and thus drives a wedge between the benchmark of direct supervision (where the signal is public and *costless*) and the cases of non-collusive and collusive supervision, where the supervisor demands a net-payment. Still, collusion has no bite, since any SCF that can be implemented under non-collusive supervision (including the respective payments to the supervisor) is also implementable under collusive supervision.

6.2 Timing of Collusion

We model collusion as a coordination of strategies used in the grand mechanism. This grand mechanism, however, consists of two stages: first a deterministic contract is selected and confidentially revealed to the supervisor, second agent and supervisor simultaneously send their reports. Our timing assumes that collusion takes place between these two stages. This implicitly assumes the mechanism reveals the information about the selected contract before players have a chance to meet and engage in collusion. For instance, it is only when the supervisor is sent to the agent in order to gather evidence that the two can collude, but at this time the supervisor already knows her mission, and in particular all details about her remuneration.

A conceivable alternative is to reveal the selected contract already *before* the supervisor accepts the mechanism. With mandatory participation this variant is equivalent to our main analysis, hence all results remain valid. Matters are slightly more complicated under voluntary participation. Assume an outside option for the supervisor of $\bar{v} \leq 0$, and the following timing. First, a deterministic contract γ is selected from Γ according to q. This choice is confidentially revealed to the supervisor who then decides whether or not to participate. The agent decides whether to participate without receiving further information. If both participate, the third party offers a side mechanisms and reporting proceeds as in our main analysis. The crucial difference is that we have to ensure the supervisor's participation for *every* deterministic contract from the menu Γ, and not only in expectation. Hence, the supervisor's expected wage in every deterministic contract has to exceed the value of her outside option. Recall from the proof of Proposition 1 that all contracts $\gamma_1, \ldots, \gamma_m$ are constructed such that they pay weakly positive wages. Wages are negative only in contract

γ_0. Hence, the supervisor participates for *all* deterministic contracts from the menu Γ if and only if she participates in contract γ_0.

Furthermore, in our construction contract γ_0 pays a deterministic wage. Increasing the probability of contract γ_0 lowers, in absolute terms, the (negative) wage paid in this contract. Moreover, this wage converges to zero as $q(\gamma_0)$ converges to one, though it always remains strictly negative. As a consequence, we find a one-sided randomized menu that implements $(\mathbf{x}, \mathbf{t})$ without paying the supervisor in our adjusted timing, whenever the supervisor's outside option is strictly negative. When $\bar{v} = 0$ we cannot use negative wages but implementation remains possible with (net-)payments to the supervisor which vanish in the limit, i.e. are *virtually* zero.

Proposition 2. *Consider the timing where the supervisor learns the selected deterministic contract before her participation decision, and has outside option $\bar{v}$. If $\bar{v} < 0$, Proposition 1 continues to hold. If $\bar{v} = 0$, there exists for any $\varepsilon > 0$ a one-sided randomized menu with properties (i) and (ii) of Proposition 1, and*

(iii′) $\sum_i q(\gamma_i) w_i(\theta, \tau, \tau) < \varepsilon$, and $w_i(\theta, \tau, \tau) \geq 0$ for all $(\theta, \tau) \in \Theta \times \mathcal{T}$.

In the proof of Proposition 2, for the case of $\bar{v} = 0$, we set all wages in contract γ_0 to zero, and thus satisfy the supervisor's participation constraint for every deterministic contract from the menu Γ. From her bonus payments in contracts $\gamma_1, \ldots, \gamma_m$, the supervisor now earns a strictly positive wage in expectation over the entire menu Γ. But recall these bonus payments under $\gamma_1, \ldots, \gamma_m$ are $\Delta + \kappa$, and thus independent of the distribution q. Now let the probabilities for these contracts converge to zero, i.e. let the probability of contract γ_0 converge to one. This yields implementation with – in the limit – vanishing wage payments.

6.3 Preferences of the Supervisor

The literature on collusive supervision often assumes specific preferences for the supervisor which depart from risk-neutrality.[22] Any departure from risk-neutrality exacerbates the implementation problem, because it makes lotteries costly, and/or prohibits (large) penalties.

[22] One reason is the common, but false, perception that collusion can easily be overcome in the case of a risk neutral supervisor, as discussed for instance by Faure-Grimaud et al. (2003). This result, however, hinges on the restrictions to i) a binary type space and ii) a single-crossing property of the agent's preferences. If we dismiss either assumption, even a risk neutral supervisor imposes severe limits on implementability: for the case of i) a larger type space, we refer to Celik (2009) who considers three types, and for the case of ii) preferences violating single crossing, see Lemma 3.

Nevertheless, one-sided randomized menus continue to mitigate the problem of collusion even under a wide range of commonly used preference functions for the supervisor.

6.3.1 Limited Liability

We fist consider limited liability, where the supervisor's utility function is given by

$$U_S(x, w, \theta) = \begin{cases} w, & w \geq \bar{w}, \\ -\infty, & w < \bar{w}, \end{cases} \tag{9}$$

for some $\bar{w} \leq 0$.[23] When the supervisor is subject to limited liability, contracts cannot use arbitrarily large negative wages. Similar to our discussion in the previous section, this is problematic only regarding contract γ_0 from our construction. As long as $\bar{w} < 0$, i.e. wages can be negative (though not too negative), modifying the distribution q on Γ allows for increasing all wages beyond $\bar{w}$, and thus yields implementability. When $\bar{w} = 0$, negative wages are ruled out per se, leaving no room for balancing the (strictly positive) bonus payments to the supervisor. Hence, implementation without paying the supervisor is not possible, but the supervisor's expected wage can be made arbitrarily small.

Proposition 3. *Suppose the supervisor is subject to limited liability. If $\bar{w} < 0$, Proposition 1 continues to hold. If $\bar{w} = 0$, the second part of Proposition 2 applies, i.e. any SCF $(\mathbf{x}, \mathbf{t})$ that is implementable under direct supervision can be implemented under collusive supervision with vanishing wages.*

Effectively, limited liability imposes constraints which resemble participation constraints as well as the adjusted timing of the mechanisms as studied in the previous section. Hence, the proof of Proposition 3 mirrors that of Proposition 2.

6.3.2 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

[23] As before, we rule out $\bar{w} > 0$. 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.

satisfies $v(0) = 0$. The latter assumption is a normalization to stay in line with the direct supervision benchmark.

In our baseline model, i.e. with mandatory participation, risk aversion imposes no further restriction. The same mechanism used to prove Proposition 1 implements $(\mathbf{x}, \mathbf{t})$ even with a risk averse supervisor and without paying any net wage. Note that *after* revealing the selected deterministic contract to the supervisor, her outcome from every possible report is deterministic, and hence reporting incentives are not affected by risk aversion. However, the supervisor's expected utility is strictly negative. She is subject to risk stemming from the uncertainty in (Γ, q). Her expected wage is zero, hence her expected utility is strictly negative by strict concavity of $v(\cdot)$.[24]

With both risk aversion and voluntary participation matters become more intricate. As long as the outside option is strictly negative it is possible to design the mechanisms such that there is only little randomness and hence risk preferences do not prevent implementation. When the supervisor's outside option is zero and she is risk-averse, implementation without net payments becomes impossible, but implementation with vanishing wages is still possible.

Proposition 4. *Suppose the supervisor is risk averse as specified above and has an outside option $\bar{v} \leq 0$. If $\bar{v} < 0$, Proposition 1 continues to hold. If $\bar{v} = 0$, the second part of Proposition 2 applies, i.e. any SCF $(\mathbf{x}, \mathbf{t})$ that is implementable under direct supervision can be implemented under collusive supervision with vanishing wages.*

6.4 Preferences of the Agent

Regarding the preferences of the agent our model is less flexible. The agent's transfer in a one-sided randomized menu is stochastic, hence incentives are directly affected when the agent is risk-averse with respect to monetary payments. Implementing a specific SCF $(\mathbf{x}, \mathbf{t})$ therefore becomes more problematic. From the perspective of a principal, who designs the mechanism to implement a specific choice rule $\mathbf{x}$, we can nevertheless show that using one-sided randomized menus allows for strictly reducing (expected) transfer payments as compared to using deterministic contracts.

With regards to limited liability, our results remain unaffected when the limit is not too

[24]Note that we assume the supervisor cares about the sum of monetary transfers stemming from the wage paid by the grand mechanism and side payments. Hence, the supervisor's incentives basically coincide with the risk-neutral case, since $v(\cdot)$ is strictly increasing.

tight. The mechanism used for proving Proposition 1 uses penalties for the agent, but these are not arbitrarily large and can hence still be used. Furthermore, our proof uses *uniform* penalties simultaneously deterring *all* deviations. That is, we can alter these penalties in two ways to meet limited liability constraints. First, we can use more nuanced deterrents depending on the realized deterministic contract and reported state rather than the uniform structure currently used, see eq. (12) in the Appendix. This allows targeting every potential deviation separately, thereby requiring fewer and lower 'penalties' to the agent. Second, we currently impose penalties even for reports which the agent would never unilaterally deviate to (such as, in a single-crossing scenario, a high-cost type claiming to have low costs). Here, again, a more subtle use of penalties in the presence of limited liability constraints is possible. Furthermore, limiting the range of transfers to the agent still leaves enough leeway to implement the SCF $(\mathbf{x}, \mathbf{t})$ with fewer wage payments to the supervisor as compared to only using deterministic contracts. In particular, it is still true that implementation may completely fail with deterministic contracts (as in our introductory example) but is possible with some wage payments to the supervisor when using one-sided randomized menus. A general answer depends on the specific assumptions imposed, regarding liability limits, participation constraints as well as the agent's utility function $u(x, \theta)$. We leave this topic for future research.

6.4.1 An Interested Supervisor

So far we have assumed that the supervisor's utility only depends on the monetary wage w she obtains. That is, she did not have any preferences over the alternative x chosen, and her preferences do not depend on the information state (θ, τ). The only reason for the supervisor to forge a report therefore was stemming from the possibility of collusion, providing her with either an increased payment from the grand mechanism or a bribe from the side mechanism. In the case of an interested supervisor, and particularly with preferences over the chosen alternative, the problem of additional incentives to misreport arises. The supervisor may now even find unilateral deviations beneficial since these may cause the implementation of more preferred alternatives, even if at the cost of a decreased wage.

To prevent such unilateral deviations, we can extend the logic of our one-sided randomized menu to *two* sides. That is, we also endow the agent with payoff-relevant private information vis-à-vis the supervisor. This way, the supervisor is uncertain about the value of both

her unilateral deviation and possible collusive coordination. Effectively, the randomization over a menu of deterministic contracts is revealed to agent and supervisor along different dimensions. Again, we leave a more detailed analysis to future research.

6.5 Supervisory Information

As a final robustness check, we analyze three alternative assumptions regarding supervisory information.

Timing of Information. In many applications the signal is observed only after entering a contract. For instance auditors are first hired (i.e. 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. That is, asymmetric information of agent and supervisor vis-à-vis the mechanism designer arrives only *post*contractually. With such an assumption all our results go through, many even get simpler to achieve, because the supervisor's participation constraint is weaker.

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 $E(\tau) \subseteq \mathcal{T}$.[25] Effectively there are fewer deviations to consider, and thus implementation becomes simpler to achieve.

Hard Information. Most work on collusive supervision assume evidence is hard. Formally, the signal observed by the supervisor either conveys the agent's true type θ or the supervisor does not find evidence at all, i.e. the signal is $\emptyset$.[26] The report can then be to reveal the evidence, or to claim not having received any. 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 all our results go through also for this case.

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

[26] See for instance Kofman and Lawarrée (1996*b*), Kessler (2004), as well as Burlando and Motta (2015).

6.6 Commitment

When stochastic mechanisms are used, the literature generally assumes the designer's ability to commit to a distribution from which the outcome is drawn. Analogously, we have assumed that a designer using one-sided randomized menus can commit to a distribution over the menu of contracts. The only difference lies in the supervisor being privately informed about the realized contract.

The mechanism designer (or principal) may not be indifferent between the different outcomes in the support of a stochastic mechanism. The same holds true for a one-sided randomized menu, where the designer may prefer some contract realizations over others. In the introductory example of Section 3, for example, the firm prefers contracts γ_1 and γ_2 over γ_0 since the latter yields a large compensation to the manager at only small gains extracted from the supervisor.

However, we can modify this mechanism such that the firm indeed is indifferent between the contracts in the menu. Consider an increased wage $\widehat{w}_i = w_i + \frac{33}{2}$ to the supervisor under contracts γ_1 and γ_2, irrespective of the report profile. This shift affects neither the manager's nor the supervisor's optimal reporting strategies and does not open up more room for side contracts. Hence, it only changes the equilibrium outcome in that the firm's expected total payments are now equal to 11 in *each* of the three contracts. Consequently, the firm is willing to randomize between the three contracts as prescribed by $q(\cdot)$.

A similar logic can be applied to more general one-sided randomized menus. Recall the construction of contracts used in the proof of Proposition 1: not knowing the realized signal τ, all contracts $\gamma_1, \dots, \gamma_m$ are equivalent in expected revenue for the mechanism designer. Only γ_0 differs in that it extracts a payment from the supervisor and pays a compensation to the agent. In order do without a commitment assumption, we hence have to match net payments between these contracts. (Note that in equilibrium, all contracts in the menu implement the same alternative x and we can hence ignore the designer's preferences over alternatives.) Contract γ_0 was constructed such that the supervisor's payoff is independent of the report profile $(\theta^a, \tau^a, \tau^s)$. For all other contracts, her payoff depends on the report prescribed by the side mechanism and hence on her input to this side mechanism. We now increase her wage for all contracts $\gamma_1, \dots, \gamma_m$ independently of the report by a constant such that $\mathbb{E}[t_l(\theta, \tau) + w_l(\theta, \tau)] = \mathbb{E}[t_0(\theta, \tau) + w_0(\theta, \tau)]$ $\forall l \in \{1, \dots, m\}$. In this way, we align the

designer's net payments across contracts, and thus achieve indifference. Note that adding a constant to the supervisor's wage affects her reporting behavior neither in the grand mechanism nor in the side mechanism. The new mechanism implements $(\mathbf{x}, \mathbf{t})$, though with positive expected wage to the supervisor. Consequently, there are no implementation issues even under the relaxed commitment assumption. Whether or not there is a need for paying the supervisor, i.e. whether there are collusive rents, remains an open question since there may be other beneficial mechanisms in this scenario.

7 Conclusion

This paper studies the issue of collusion when a supervisor is employed to extract the information held by an agent. We use a mechanism design framework to analyze the constraints which collusion puts on, first, the set of social choice functions which are implementable and, second, the costs caused by rent payments to the colluding parties. Previous literature almost exclusively focused on direct mechanisms (or a subset thereof) and came to the conclusion that collusion impedes both implementability and rent extraction. We show that neither result persists if more general mechanisms are used. In particular, we introduce *one-sided randomized* menus which provide the supervisor with payoff relevant private information vis-à-vis the agent. We thus endogenously create an informational asymmetry between the colluding parties. Modeling their side agreements as a form of bilateral trade, the introduction of such asymmetric information can cause a complete breakdown of their bargaining. That is, no side mechanism can be found which departs from the non-cooperative equilibrium and at the same time is incentive compatible, individual rational and ex-ante budget balanced. When haggling over adequate bribes, the agent does not know whether he faces a 'lemons' supervisor in the spirit of Akerlof (1970), who is cheap to collude with but of little value for the agent. Our main result is that any social choice function which can be implemented under direct supervision – that is, when the supervisory information is public – is also implementable under collusive supervision. In addition, extracting the supervisory signal is costless.

Our results prove robust to various model changes, in particular common assumptions on the supervisor's utility function, such as risk aversion, limited liability or voluntary participation. For each of these scenarios, the social choice functions remain at least virtually

implementable and at costs which vanish in the limit. It remains an open question, however, whether the results also extend to the case of two-sided asymmetric information: while in our model, ex ante information is nested along the hierarchy (i.e. the agent holds private information vis-à-vis the supervisor, but not vice versa), there are situations where each of the colluding party holds private information. This is particularly relevant in an auction setting of two or more agents. It may well be possible to exploit the introduction of endogenous asymmetric information in a way similar to the one presented in this paper, i.e. providing *each* of the agents with additional payoff-relevant information he holds privately. In addition, in our setting the supervisor's preferences are independent of the information thought to be extracted – that is, the agent's type – and of the alternative chosen by the mechanism. In a more general framework, each of the agents could have preferences depending on the information state and the chosen alternative. Again, multi-sided randomized menus could be used to extract each agent's privately held information by injecting bargaining frictions into the coalition. There may, however, well be other mechanism sharing the benefits of the ones we propose. As the focus of this paper is on introducing a specific methodology and its potential benefits, we leave the exploration of this topic for future research.

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. (2018), 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.

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.

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.

Kessler, A. S. (2004), 'Optimal auditing in hierarchical relationships', *Journal of Institutional and Theoretical Economics* **160**(2), 210–231.

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

Kofman, F. and Lawarrée, J. (1996*a*), 'On the optimality of allowing collusion', *Journal of Public Economics* **61**(3), 383–407.

Kofman, F. and Lawarrée, J. (1996*b*), '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.

Moore, J. and Repullo, R. (1990), 'Nash implementation: a full characterization', *Econometrica: Journal of the Econometric Society* pp. 1083–1099.

Ortner, J. and Chassang, S. (2018), 'Making collusion hard: Asymmetric information as a counter-corruption measure', *Journal of Political Economy* (forthcoming).

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.

Samuelson, W. (1984), 'Bargaining under asymmetric information', *Econometrica* **52**(4), 995–1005.

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

Proof of Lemma 1. Follows directly from invoking the 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 Lemma 2. Consider a social choice function $(\mathbf{x},\mathbf{t})$ which is implementable under direct supervision. Let $\widetilde{w} < 0$, $\widetilde{x} \in \mathcal{X}$, and $\widetilde{t} < \min_{\theta,\tau} u(\mathbf{x}(\theta,\tau),\theta) + \mathbf{t}(\theta,\tau) - u(\widetilde{x},\theta)$. Define a deterministic contract $\gamma = (x,t,w)$ by

$$(x,t,w)(\theta,\tau,\tau') = \begin{cases} (\mathbf{x}(\theta,\tau), \mathbf{t}(\theta,\tau), 0), & \text{if } \tau = \tau', \theta \in \Theta(\tau), \\ (\widetilde{x},\widetilde{t},\widetilde{w}), & \text{else.} \end{cases}$$

The deterministic contract γ implements the SCF $(\mathbf{x},\mathbf{t})$ if it exhibits an equilibrium where both the agent and the supervisor report their information truthfully. Conditional on the agent reporting truthfully, the supervisor prefers to do so as well since $\widetilde{w} < 0$. By the definition of $\widetilde{t}$ the agent does not want to unilaterally deviate to a non-conforming signal-report. Implementability of $(\mathbf{x},\mathbf{t})$ under direct supervision further implies the agent

has no incentive to report any $\theta' \in \Theta(\tau)$ different from his true type θ. Lastly, sending report $\theta' \notin \Theta(\tau)$, while reporting τ truthfully, is not optimal by the definition of $\widetilde{t}$. $\square$

Proof of Proposition 1. Fix a social choice function $(\mathbf{x}, \mathbf{t})$ that can be implemented under direct supervision. For any $(\theta, \tau) \in \Theta \times \mathcal{T}$, and all $1 \leq l \leq m$ and $1 \leq k \leq n$ define

$$\Delta_{kl}(\theta, \tau) := u(\mathbf{x}(\theta_k, \tau_l), \theta) + t(\theta_k, \tau_l) - u(\mathbf{x}(\theta, \tau), \theta) - t(\theta, \tau), \tag{10}$$

and define $\Delta := \max_{l,k,\theta,\tau} \Delta_{lk}(\theta, \tau)$. Assume without loss of generality $\Delta > 0$, since otherwise the desired allocation was implementable without using the supervisor.

We next define a one-sided randomized menu (Γ, q) with $\Gamma = \{\gamma_0, \gamma_1, \ldots, \gamma_m\}$. A simple contract γ_i asks the agent to report a type $\theta^a \in \Theta$ and a signal $\tau^a \in \mathcal{T}$, and the supervisor to report a signal $\tau^s \in \mathcal{T}$. Contract γ_l specifies for each report vector $(\theta^a, \tau^a, \tau^s) \in \Theta \times \mathcal{T} \times \mathcal{T}$ an allocation $x_l(\theta^a, \tau^a, \tau^s)$, a transfer to the agent $t_l(\theta^a, \tau^a, \tau^s)$ and a wage to the supervisor $w_l(\theta^a, \tau^a, \tau^s)$. For $1 \leq l \leq m$ define the simple contract γ_m as follows:

$$x_l(\theta^a, \tau^a, \tau^s) = \begin{cases} \mathbf{x}(\theta^a, \tau^a), & \tau^a = \tau^s, \theta^a \in \Theta(\tau^a), \\ \widetilde{x}, & \text{otherwise}, \end{cases} \tag{11}$$

and

$$t_l(\theta^a, \tau^a, \tau^s) = \begin{cases} \mathbf{t}(\theta^a, \tau^a) - y\Delta, & \tau^a = \tau^s = \tau_l, \theta^a \in \Theta(\tau_l), \\ \mathbf{t}(\theta^a, \tau^a) & \tau^a = \tau^s \neq \tau_l, \theta^a \in \Theta(\tau^a), \\ -\chi, & \text{otherwise}, \end{cases} \tag{12}$$

as well as

$$w_l(\theta^a, \tau^a, \tau^s) = \begin{cases} \Delta + \kappa, & \tau^a = \tau^s = \tau_l, \\ 0, & \tau^a = \tau^s \neq \tau_m \\ 0, & \text{otherwise}. \end{cases} \tag{13}$$

Define the simple contract γ_0 as follows:

$$x_0(\theta^a, \tau^a, \tau^s) = \begin{cases} \mathbf{x}(\theta^a, \tau^a), & \tau^a = \tau^s, \theta^a \in \Theta(\tau^a), \\ \widetilde{x}, & \text{otherwise}, \end{cases} \tag{14}$$

and

$$t_0(\theta^a, \tau^a, \tau^s) = \begin{cases} \mathbf{t}(\theta^a, \tau^a) + \frac{\delta}{1-\delta}\frac{y\Delta}{m}, & \tau^a = \tau^s, \theta^a \in \Theta(\tau^a), \\ -\chi, & \text{otherwise}, \end{cases} \tag{15}$$

as well as

$$w_0(\theta^a, \tau^a, \tau^s) = -w_0 \tag{16}$$

for a fixed $w_0 \geq 0$. Further assume $\widetilde{x} \in X$, $\kappa > 0$, $\delta \in (0,1)$, and

$$y > \max\left\{\tfrac{m}{\delta} - 1, m + \tfrac{\kappa}{\Delta}\right\}, \tag{17}$$

$$\chi > u(\widetilde{x}, \theta_i) - u(\mathbf{x}(\theta_i, \tau_j), \theta_i) - t(\theta_i, \tau_j), \qquad \forall 1 \leq i \leq n, 1 \leq j \leq m. \tag{18}$$

Finally, let $q(\gamma_0) = 1 - \delta$, and $q(\gamma_i) = \delta/m$ for $1 \leq i \leq m$.

We now prove a more general statement, allowing us to later on easily incorporate settings where we have additional constraints such as limited liability, risk aversion or interim participation (see Propositions 2 and 3). These settings will put a lower bound on the parameter w_0. Proposition 1 will follow from this statement.

Proposition A.1. *Given an SCF* $(\mathbf{x}, \mathbf{t})$*, consider the one-sided randomized menu* (Γ, q) *defined by eq.* (11) *to eq.* (18)*.*

i) $\forall w_0 > 0$ $\exists \delta \in (0,1)$ *such that* (Γ, q) *implements* $(\mathbf{x}, \mathbf{t})$ *without net payments to the supervisor,*

ii) If $w_0 = 0$*,* (Γ, q) *virtually implements* $(\mathbf{x}, \mathbf{t})$ *without net payments to the supervisor for* $\delta \to 0$*, i.e.* $\forall \epsilon > 0$ *we have* $\sum_i q(\gamma_i) w_i(\theta, \tau, \tau) < \varepsilon$ *for all* $(\theta, \tau) \in \Theta \times \mathcal{T}$*.*

Proof of Proposition A.1. For strictly positive w_0, consider

$$\delta = \frac{m w_0}{m w_0 + \Delta + \kappa}. \tag{19}$$

Since $w_0, \kappa > 0$ by assumption and $\Delta \geq 0$, we know that $\delta \in (0,1)$ as required. Note that given the (Γ, q) is constructed such that, *if* agent and supervisor report truthfully, the supervisor is indeed not paid in expectation: her expected payoff (prior to learning which simple contract was drawn from the menu) is given by $\frac{\delta}{m}(\Delta + \kappa) - (1 - \delta) w_0 = 0$. We therefore have to show that (Γ, q) exhibits an equilibrium with truthful reports.

First consider unilateral deviations. Obviously, (Γ, q) features a non-cooperative equilibrium where the agent reports his type truthfully, and the supervisor reports the signal truthfully. The latter receives a bonus only if both reports on the signal coincide, hence *given* that the agent's report is truthful the supervisor prefers to do so as well. Next consider the agent. The payment $-\chi$ deters misreports of the signal, provided the supervisor reports the true signal. Incentive compatibility of the menu $\big(\mathbf{x}(\cdot,\tau),\mathbf{t}(\cdot,\tau)\big)$ for all τ implies the agent also prefers to report his type truthfully.

In the remainder we show there is no feasible side-agreement impeding the non-cooperative equilibrium. A side mechanism (w.l.o.g. direct) consists of a collection of probabilities p_{kl}^{ij}, where $i \in \{1,\dots,n\}$ is the agent's report to the side-mechanism, $j \in \{0,1,\dots,m\}$ is the supervisor's report to the side mechanism, and (θ_k,τ_l,τ_l) is the collective report to the grand mechanism. Furthermore, there are probabilities $p_{\emptyset}^{ij}$ for sending non-conforming signal reports, where we do not have to distinguish exactly which individual reports are sent, and which type the agent reports, as our grand mechanism does not distinguish either. Furthermore the side mechanism specifies payments (b_{ij}^a, b_{ij}^s), where b_{ij}^a (and b_{ij}^s, respectively) is the payment the agent (supervisor) receives after type report θ_i by the agent and a simple contract report γ_j by the supervisor. Recall we can invoke a revelation principle on the collusion stage, i.e. we can focus on side mechanisms inducing truthtelling of both agent and supervisor. If the agent truthfully reports the true type, the supervisor's expected bribe given a signal τ_d and his report γ_j is given by

$$\mathbf{b}_j^s = \sum_{i=1}^{n} \pi_i^d b_{ij}^s, \tag{20}$$

where we omit the true signal τ_d in the shorthand notation since we consider a fixed signal in the following. The agent's expected bribe given truthtelling of the supervisor and his report θ_i is similarly defined by

$$\mathbf{b}_i^a = \sum_{j=1}^{m} q(\gamma_j) b_{ij}^a. \tag{21}$$

In the following fix a true signal $\tau_d \in \mathcal{T}$ observed by both the agent and the supervisor. As private information, the agent knows his type θ, and the supervisor knows the simple contract γ. Recall, π_i^d is the conditional probability that the agent has type θ_i when the

signal τ_d relized. Define

$$\mathbf{p}_l^j = \sum_{i=1}^{n}\sum_{k=1}^{n} \pi_i^d p_{kl}^{ij}, \tag{22}$$

the probability that the side-mechanism triggers report τ_l when the supervisor reports γ_j, and

$$\mathbf{p}_\emptyset^j = \sum_{i=1}^{n} \pi_i^d p_\emptyset^{ij}, \tag{23}$$

the probability that the side-mechanism triggers a non-conforming signal-report when the supervisor reports γ_j.

The supervisor, who knows that simple contract γ_d was selected, participates in the side mechanism, whenever $\mathbf{p}_d^d(\Delta+\kappa) + \mathbf{b}_d^s \geq \Delta + \kappa$. Hence, in any incentive compatible and individually rational side mechanism we must have

$$\mathbf{b}_d^s \geq (1 - \mathbf{p}_d^d)(\Delta + \kappa). \tag{24}$$

The following Lemma provides some necessary conditions for an incentive compatible side mechanism.

Lemma A.1. *If the side-mechanism is incentive compatible, then*

$$\mathbf{p}_j^j \geq \mathbf{p}_j^0 \qquad \forall j = 1, \ldots, m, \tag{25}$$

$$\mathbf{p}_j^j + \mathbf{p}_k^k \geq \mathbf{p}_j^k + \mathbf{p}_k^j \qquad \forall j, k = 1, \ldots, m. \tag{26}$$

Proof. The supervisor of type $j = 1, \ldots, m$ prefers reporting truthfully, whenever

$$\mathbf{p}_j^j(\Delta+\kappa) + \mathbf{b}_j^s \geq \mathbf{p}_j^k(\Delta+\kappa) + \mathbf{b}_k^s, \quad \forall\ 0 \leq k \leq m. \tag{27}$$

Similarly, the supervisor of type 0 prefers reporting truthfully, whenever

$$\mathbf{b}_0^s \geq \mathbf{b}_j^s, \quad \forall\ 1 \leq j \leq m. \tag{28}$$

Adding (27) and (28) yields $\mathbf{p}_j^j(\Delta+\kappa) + \mathbf{b}_j^s + \mathbf{b}_0^s \geq \mathbf{p}_j^0(\Delta+\kappa) + \mathbf{b}_0^s + \mathbf{b}_j^s$, and thus (25). Similarly, adding (27) for $j \to k$ and $k \to j$ yields

$$(\mathbf{p}_j^j + \mathbf{p}_k^k)(\Delta+\kappa) + \mathbf{b}_j^s + \mathbf{b}_k^s \geq (\mathbf{p}_j^k + \mathbf{p}_k^j)(\Delta+\kappa) + \mathbf{b}_k^s + \mathbf{b}_j^s,$$

and thus (26). □

We continue the proof of Proposition A.1 i). From the supervisor's incentive compatibility constraints (27) and (28) we further get for all $j \neq d, 0$

$$\mathbf{b}_j^s \geq \mathbf{b}_0^s + (\mathbf{p}_j^0 - \mathbf{p}_j^j)(\Delta + \kappa) \geq \mathbf{b}_d^s + (\mathbf{p}_j^0 - \mathbf{p}_j^j)(\Delta + \kappa). \tag{29}$$

Hence,

$$\begin{aligned} \mathbb{E}(\mathbf{b}^S) = \sum_{j=1}^{m} \frac{\delta}{m} \mathbf{b}_j^S + (1-\delta) b_0^S &\geq \mathbf{b}_d^s + \sum_{j \neq d} \frac{\delta}{m} \left(\mathbf{p}_j^0 - \mathbf{p}_j^j\right)(\Delta + \kappa) \\ &\geq \left(1 - \mathbf{p}_d^d\right)(\Delta + \kappa) + \frac{\delta}{m}(\Delta + \kappa) \sum_{j \neq d} \left(\mathbf{p}_j^0 - \mathbf{p}_j^j\right). \end{aligned} \tag{30}$$

Next consider the agent. When of type θ_i, the agent's expected payoff from participating in the side-mechanism is

$$\begin{aligned} \mathbf{b}_i^a \quad &+ \quad \sum_{j=1}^{m} \frac{\delta}{m} \left(\sum_{k=1}^{n} \sum_{l=1}^{m} p_{kl}^{ij} \Big[u(x_j(\theta_k, \tau_l, \tau_l), \theta_i) + t_j(\theta_k, \tau_l, \tau_l) \Big] \right) \\ &+ \quad (1-\delta) \sum_{k=1}^{n} \sum_{l=1}^{m} p_{kl}^{i0} \Big[u(x_0(\theta_k, \tau_l, \tau_l), \theta_i) + t_0(\theta_k, \tau_l, \tau_l) \Big] \\ &+ \quad \left\{ \frac{\delta}{m} \sum_{j=1}^{m} p_{\emptyset}^{ij} + (1-\delta) p_{\emptyset}^{i0} \right\} (u(\widetilde{x}, \theta_i) - \chi) \end{aligned}$$

As explained above, the agent's expected payoff is $u(\mathbf{x}(\theta_i, \tau_d), \theta_i) + \mathbf{t}(\theta_i, \tau_d)$ from playing non-coperatively, that is from refusing to participate in the side mechanism. He thus

participates in the side mechanism, whenever

$$
\begin{aligned}
\mathbf{b}_i^a \;\geq\;& \sum_{j=1}^{m}\frac{\delta}{m}\sum_{k=1}^{n}\sum_{l=1}^{m}p_{kl}^{ij}\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)-u(x_j(\theta_k,\tau_l,\tau_l),\theta_i)-t_j(\theta_k,\tau_l,\tau_l)\Big]\\
&+(1-\delta)\sum_{k=1}^{n}\sum_{l=1}^{m}p_{kl}^{i0}\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)-u(x_0(\theta_k,\tau_l,\tau_l),\theta_i)-t_0(\theta_k,\tau_l,\tau_l)\Big]\\
&+\left\{\frac{\delta}{m}\sum_{j=1}^{m}p_{\emptyset}^{ij}+(1-\delta)p_{\emptyset}^{i0}\right\}\Big[u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)-u(\widetilde{x},\theta_i)+\chi\Big]\\
\geq\;& -\frac{\delta}{m}\sum_{j=1}^{m}\left(\sum_{k=1}^{n}\sum_{l\neq d}p_{kl}^{ij}\Delta-\sum_{k=1}^{n}p_{kj}^{ij}y\Delta\right)-(1-\delta)\left(\sum_{k=1}^{n}\sum_{l\neq d}p_{kl}^{i0}\Delta+\frac{\delta}{1-\delta}\frac{y\Delta}{m}\right)\\
&+\left\{\frac{\delta}{m}\sum_{j=1}^{m}p_{\emptyset}^{ij}+(1-\delta)p_{\emptyset}^{i0}\right\}\bar{\Delta}\\
=\;& -\delta\frac{\Delta}{m}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l\neq d}p_{kl}^{ij}-(1-\delta)\Delta\sum_{k=1}^{n}\sum_{l\neq d}p_{kl}^{i0}+\delta\frac{y\Delta}{m}\sum_{j=1}^{m}\sum_{k=1}^{n}\left(p_{kj}^{ij}-p_{kj}^{i0}\right)\\
&+\left\{\frac{\delta}{m}\sum_{j=1}^{m}p_{\emptyset}^{ij}+(1-\delta)p_{\emptyset}^{i0}\right\}\bar{\Delta},
\end{aligned}
$$

where the second inequality uses

(i) $u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)\geq\max\{u(\mathbf{x}(\theta_k,\tau_d),\theta_i)+\mathbf{t}(\theta_k,\tau_d),u(\widetilde{x},\theta_i)-\chi\}$ for all $\theta_k\in\Theta(\tau_d)$, by incentive compatibility of $(\mathbf{x},\mathbf{t})$ and the definitions of $\widetilde{x}$ and χ,

(ii) $u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)-\max\{u(\mathbf{x}(\theta_k,\tau_l),\theta_i)+\mathbf{t}(\theta_k,\tau_l),u(\widetilde{x},\theta_i)-\chi\}\geq-\Delta$ for all τ_l and $\theta_k\in\Theta(\tau_l)$ by the definitions of Δ,κ and $\widetilde{x}$,

(iii) $\bar{\Delta}:=\min_{i,d}\bar{\Delta}(\theta_i,\tau_d)=\min_{i,d}\chi+u(\mathbf{x}(\theta_i,\tau_d),\theta_i)+\mathbf{t}(\theta_i,\tau_d)-u(\widetilde{x},\theta_i)>0$ by (18).

Summing the lower bounds for agent's payment in the side-mechanism yields

$$
\begin{aligned}
\mathbb{E}(\mathbf{b}^A) &= \sum_{i=1}^{n} \pi_i^d b_i^a \\
&\geq -\delta\frac{\Delta}{m}\sum_{j=1}^{m}\sum_{l\neq d}\mathbf{p}_l^j - (1-\delta)\Delta\sum_{l\neq d}\mathbf{p}_l^0 + \delta\frac{\Delta y}{m}\sum_{j=1}^{m}(\mathbf{p}_j^j - \mathbf{p}_j^0) \\
&\quad + \left\{\frac{\delta}{m}\sum_{j=1}^{m}\mathbf{p}_\emptyset^j + (1-\delta)\mathbf{p}_\emptyset^0\right\}\bar{\Delta} \\
&= \delta\frac{\Delta y}{m}\sum_{j=1}^{m}(\mathbf{p}_j^j - \mathbf{p}_j^0) + \left\{\frac{\delta}{m}\sum_{j=1}^{m}\mathbf{p}_\emptyset^j + (1-\delta)\mathbf{p}_\emptyset^0\right\}\bar{\Delta} \\
&\quad -\Delta\left\{\frac{\delta}{m}\sum_{l\neq d}\mathbf{p}_l^d + \frac{\delta}{m}\sum_{j>l\neq d}(\mathbf{p}_l^j + \mathbf{p}_j^l) + \frac{\delta}{m}\sum_{j\neq d}\mathbf{p}_j^j + (1-\delta)\sum_{l\neq d}\mathbf{p}_l^0\right\} \\
&\overset{(26)}{\geq} \delta\frac{\Delta y}{m}\sum_{j=1}^{m}(\mathbf{p}_j^j - \mathbf{p}_j^0) + \left\{\frac{\delta}{m}\sum_{j=1}^{m}\mathbf{p}_\emptyset^j + (1-\delta)\mathbf{p}_\emptyset^0\right\}\bar{\Delta} \\
&\quad -\Delta\left\{\frac{\delta}{m}\sum_{l\neq d}\mathbf{p}_l^d + \delta\frac{m-1}{m}\sum_{j\neq d}\mathbf{p}_j^j + (1-\delta)\sum_{l\neq d}\mathbf{p}_l^0\right\} \\
&= \delta\frac{\Delta y}{m}\sum_{j=1}^{m}(\mathbf{p}_j^j - \mathbf{p}_j^0) + \left\{\frac{\delta}{m}\sum_{j=1}^{m}\mathbf{p}_\emptyset^j + (1-\delta)\mathbf{p}_\emptyset^0\right\}\bar{\Delta} \\
&\quad -\Delta\left\{\frac{\delta}{m}(1-\mathbf{p}_d^d - \mathbf{p}_\emptyset^d) + \delta\frac{m-1}{m}\sum_{j\neq d}\left(\mathbf{p}_j^j - \mathbf{p}_j^0\right) + \left(1-\frac{\delta}{m}\right)(1-\mathbf{p}_d^0 - \mathbf{p}_\emptyset^0)\right\} \\
&= \delta\frac{\Delta y}{m}\sum_{j=1}^{m}(\mathbf{p}_j^j - \mathbf{p}_j^0) + \left\{\frac{\delta}{m}\sum_{j=1}^{m}\mathbf{p}_\emptyset^j + (1-\delta)\mathbf{p}_\emptyset^0\right\}\bar{\Delta} \\
&\quad -\Delta\left\{1 - \mathbf{p}_d^0 + \delta\frac{m-1}{m}\sum_{j\neq d}\left(\mathbf{p}_j^j - \mathbf{p}_j^0\right) - \frac{\delta}{m}(\mathbf{p}_d^d - \mathbf{p}_d^0) - \left(1-\frac{\delta}{m}\right)\mathbf{p}_\emptyset^0 - \frac{\delta}{n}\mathbf{p}_\emptyset^d\right\}.
\end{aligned}
$$

Adding the lower bounds on the supervisor's and the agent's payments, in any incentive compatible and individually rational side mechanism we have

$$
\begin{aligned}
\mathbb{E}(\mathbf{b}^S + \mathbf{b}^A) \geq &\Big(1-\mathbf{p}_d^d\Big)\kappa + \Delta\Big(\tfrac{\delta}{m}y + \tfrac{\delta}{m} - 1\Big)\Big(\mathbf{p}_d^d - \mathbf{p}_d^0\Big) + \tfrac{\delta}{m}\sum_{j\neq d}\Big(y\Delta - m\Delta - \kappa\Big)\Big(\mathbf{p}_j^j - \mathbf{p}_j^0\Big) \\
&+ \tfrac{\delta}{m}\bar{\Delta}\sum_{i\neq d}\mathbf{p}_\emptyset^i + \tfrac{\delta}{m}\Big(+\bar{\Delta} + \Delta\Big)\mathbf{p}_\emptyset^d + [(1-\delta)\bar{\Delta} + (1-\tfrac{\delta}{m})\Delta]\mathbf{p}_\emptyset^0. \qquad (31)
\end{aligned}
$$

Following (25), and assumptions (17) and (18), all terms on the right-hand side of (31) are non-negative. Hence, a side-mechanism is incentive compatible, individually rational and ex-ante balanced budget, only if the right-hand side in (31) equals zero. The latter requires

$\mathbf{p}_d^d = 1$, as well as $0 = \mathbf{p}_j^j - \mathbf{p}_j^0$ and $\mathbf{p}_\emptyset^j = 0$ for all $1 \leq j \leq m$. From $\mathbf{p}_d^d = 1$ and $\mathbf{p}_d^d - \mathbf{p}_d^0 = 1$ we get $\mathbf{p}_d^0 = 1$. Then $\mathbf{p}_j^0 = 0$ for all $j \neq d$ and thus also $\mathbf{p}_j^j = 0$ for all $j \neq d$, because $\mathbf{p}_j^j - \mathbf{p}_j^0 = 0$. Using (26) we have $0 \leq \mathbf{p}_j^k + \mathbf{p}_k^j \leq \mathbf{p}_j^j + \mathbf{p}_k^k = 0$ for all $j, k \neq d$, and thus $\mathbf{p}_k^j = 0$ for all $j, k \neq d$. Since $1 = \sum_{k=1}^m \mathbf{p}_k^j + \mathbf{p}_\emptyset^j$ for all $j = 0, 1, \ldots, m$, we thus have $\mathbf{p}_d^j = 1$ for all $j = 0, 1, \ldots, m$. We have thus shown that the only incentive-compatible, individually rational, and ex-ante balanced budget side mechanism entails

$$p_{kl}^{ij} = \begin{cases} 1, & l = d \\ 0, & \text{otherwise,} \end{cases}$$

for all $j \in \{1, \ldots, m\}$ and all $i, k \in \{1, \ldots, n\}$. In words, the side-mechanism commits both players to report the true signal. Incentive compatibility for the agent's type-report further implies the outcome of the side mechanism corresponds to the non-cooperative equilibrium outcome described above.

We thus have shown that the scf $(\mathbf{x}, \mathbf{t})$ is implementable under collusive supervision. As argued above, the supervisor's expected wage (before learning which simple contract was selected) is zero. This concludes the proof of Proposition A.1 i).

Now consider the case $w_0 = 0$. The above reasoning on individual and coalitional truthtelling still applies, and this holds for all $\delta \in (0, 1)$. Crucially, the expected transfer to the agent is $\mathbf{t}(\theta, \tau)$ for all δ. Regarding the supervisor's expected wage, have $\sum_i q(\gamma_i) w_i(\theta, \tau, \tau) = \frac{\delta}{m}(\Delta + \kappa)$ which converges to zero as $\delta \to 0$. Hence, the mechanisms implements $(\mathbf{x}, \mathbf{t})$ with vanishing wage payments. □

Continuing the proof of Proposition 1, we can now simplify the two-sided randomized menu (Γ, q): since we have no constraint on the supervisor's wage, we can set $w_0 = \Delta + \kappa$ and obtain $\delta = \frac{m}{m+1}$. All simple contracts $(\gamma_0, \ldots, \gamma_m)$ in Γ then are equally likely: $q(\gamma_l) = \frac{1}{m+1}$ $\forall 0 \leq l \leq m$. The mechanism takes the form depicted in Table 1. □

Proof of Lemma 3. We first show that, when $(\mathbf{x}, \mathbf{t})$ is implementable, then it is cyclically monotone. Invoking a collusion-proofness principle, we can restrict to deterministic contracts without collusion on the equilibrium path. Fix transfers $(t_1, \ldots, t_n)$ and wages $(w_1, \ldots, w_n)$ that implement $(\mathbf{x}, \mathbf{t})$. Let $(\theta^1, \theta^2, \ldots, \theta^k)$ be some sequence with $\theta^1 = \theta^k$. The SCF is implementable, thus in state θ^{l+1} there is no side mechanism that commits agent and supervisor to send joint report θ^l, making both at least as well off and one strictly better off.

That is, for any $b \in \mathbb{R}$ with $v(w(\theta^{l+1})) = v(w(\theta^l) + b)$ we have $u(x(\theta^{l+1}), \theta^{l+1}) + t(\theta^{l+1}) \geq u(x(\theta^l), \theta^{l+1}) + t(\theta^l) - b$. Since $v(\cdot)$ is strictly increasing, we have $w(\theta^l) + b = w(\theta^{l+1})$ and thus

$$u(x(\theta^l), \theta^{l+1}) - u(x(\theta^{l+1}), \theta^{l+1}) \leq t(\theta^{l+1}) - t(\theta^l) + w(\theta^{l+1}) - w(\theta^l)$$

Summing these inequalities from $l = 1, \dots k-1$ yields

$$\sum_{l=1}^{k-1} u(x(\theta^l), \theta^{l+1}) - u(x(\theta^{l+1}), \theta^{l+1}) \leq 0$$

Using $\theta^1 = \theta^k$, we thus have

$$\begin{aligned}
\sum_{l=1}^{k-1} u(x(\theta^l), \theta^{l+1}) - u(x(\theta^l), \theta^l) &= \sum_{l=1}^{k-1} u(x(\theta^l), \theta^{l+1}) - \sum_{l=1}^{k-1} u(f(\theta^l), \theta^l) \\
&= \sum_{l=1}^{k-1} u(x(\theta^l), \theta^{l+1}) - \sum_{l=1}^{k-1} u(x(\theta^{l+1}), \theta^{l+1}) \\
&= \sum_{l=1}^{k-1} u(x(\theta^l), \theta^{l+1}) - u(x(\theta^{l+1}), \theta^{l+1}) \\
&\leq 0,
\end{aligned}$$

which proves cyclical monotonicity.

To show sufficiency, notice that cyclical monotonicity is sufficient for implementing $(\mathbf{x}, \mathbf{t})$ in the absence of supervision. Hence, setting $w_1 = \cdots = w_n = 0$ also implements $(\mathbf{x}, \mathbf{t})$ under collusive supervision. □

Proof of Proposition 2. The proof directly follows from Proposition A.1: consider the mechanism (Γ, q) described above with the constraint that $w_0 = -\overline{v}$. For all simple contracts $\gamma_1, \dots, \gamma_m$, the supervisor's wage is non-negative and hence weakly greater than $\overline{v}$. For γ_0, she obtains just her outside option. Hence, she participates in the grand mechanism irrespective of the realized draw from the menu. When $\overline{v} = 0$ we get implementation with vanishing wages, as in part (ii) of Proposition A.1. □

Proof of Proposition 3. The proof mirrors that of Proposition 2. □

Proof of Proposition 4. The proof mirrors that of Proposition 2. □