

ISSN 1471-0498



DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

**MONITORING SUBCONTRACTING IN A SUPPLIERS'
HIERARCHY**

Michela Cella

Number 233

April 2005

Manor Road Building, Oxford OX1 3UQ

Monitoring Subcontracting in a Suppliers' Hierarchy.*

Michela Cella[†]

University of Oxford and Nuffield College

This Version: April 2005

Abstract

In this paper we study the delegation of a production process in a three-tier hierarchy. The principal contracts directly only with the supplier that produces the first input leaving him in charge of the contract for the production of the second input. We allow the principal to costlessly monitor the communication between the agents at the subcontracting stage in an attempt to save on informational rents and improve productive efficiency. We show that, if the contractor is free to choose the type of subcontract, he must be given additional incentives to acquire information about the subcontractor which will then be object of the monitoring. The monitoring is therefore much less effective than when the principal can force the contractor into choosing her preferred subcontract.

Keywords: Adverse Selection, Hierarchies, Delegation, Monitoring.

JEL classification: D20, D82, L22, L51

*This is a substantial revision of a chapter of my PhD thesis submitted at the LSE and previously circulated as "Monitoring of Delegated Contracting". I received useful comments and suggestions from Kevin Roberts, Antoine Faure-Grimaud, Leonardo Felli and Andrea Prat. I also wish to thank Michele Arslan, Anna Creti, Luca Deidda, Niko Matouschek, Paolo Ramezzana, Imran Rasul, Cecilia Testa and all the participants to the EOPP internal seminar. All remaining errors are mine.

[†]e-mail: michela.cella@economics.ox.ac.uk, Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, UK

1 Introduction.

Delegation of economic activity and subcontracting are widely observed phenomena, examples include the activity of a manager who is organizing a supplier network on behalf of the firm owner and the one of a prime contractor in procurement who is dealing with a subcontractor. Such diffusion has most likely been favoured by the high improvement in communication and the increased sophistication of the available forms of contracts.

We often observe a hierarchical structure where each level is linked to the lower one by a contract ruling one or more economic activities. Hierarchical decentralization involves gains from specialization and the reduction of information processing costs but it also brings about extra costs due to the loss of control over the lower levels of the organization.

Understanding whether the advantages of delegation outnumber the disadvantages is beyond the scope of this paper, our goal is instead to make progress in the understanding of the interactions between members of a hierarchy. We take the organizational form as given and we study how the informational structure is shaped by the actions of the players.

We focus, in fact, on how the efficiency of an organization or a network of suppliers is affected by the attempts of the top level of the hierarchy to regain control by monitoring the relationships between lower levels. We show that there is a gain in efficiency when the principal monitors, but that this gain is greatly reduced when we take into account the freedom and autonomy of the middle agent in choosing the amount of information that is exchanged at subcontracting. It should come as no surprise that the nature of the game depends on the observability of communication and that the scope of control in a multi-unit organization affects overall performance.

We study a setting of hierarchical contracting with three vertical layers and where contracting is restricted to adjacent layers. It can be viewed as a principal wanting to produce a final output using two inputs, one is produced by a prime-contractor with whom the principal deals directly while the second one is produced by a subcontractor that contracts and communicates only with the middle layer and has no contact with the principal. Both productive agents have private information about their marginal costs.

Using contract theory to study economic interactions between members of some hierarchical structure has proven to be quite fruitful despite being a relatively unsuccessful analytical framework to justify the existence of hierarchies due to the difficulty of incorporating the above mentioned benefits of delegation into a contract theory model. The problem with classical incentive theory based on the Revelation Principal and its variations is that *ceteris paribus* a centralized structure always weakly dominates a decentralized one.

We will study two optimal contracts, a grand contract between the principal and A_1 , the prime contractor, and a subcontract between A_1 and A_2 , the subcontractor.

The principal is confronted with additional incentives problems because when offering the contract to A_1 she has to give incentives to this agent to truthfully report not only his own type but also the type of the second one, which he will have learned at the subcontracting stage. There is a “cascading” of informational rents: first a rent is paid by A_1 to A_2 during subcontracting, then at the grand-contract stage this is subject to an additional mark-up due to the privacy of the contractor’s information vis-a-vis the principal regarding contracting costs and on top of this there is the “standard” informational rent paid by the principal to the first agent.

This mark-up on the subcontracting cost is precisely the cost of delegation, and the principal has to pay for it because what happens at the subcontracting stage is private information.

Monitoring the communication between the contractor and the subcontractor would allow the principal to reduce her total costs because she would obtain for free some information. More precisely she would monitor both the phases at the subcontracting stage: the offer of subcontract and the reply to it.

Through the monitoring of the offer the principal learns the type of the middle agent, who is left with no rent in any state of the world. The agent can neutralize this by making an offer to the bottom agent that is conditioned on his own type without revealing it, by offering a menu of contracts the agent delays the revelation of his type. This application of Myerson’s [1983] inscrutability principle is costless for both the contractor and the subcontractor and reinstates the asymmetry of information between the principal and the contractor regarding the latter’s type.

By monitoring the other stage of subcontracting, the reply, the principal learns the type of the bottom agent. Once again the player penalized by this activity is the contractor who loses the ability to manipulate the information about contracting costs for which, in the standard set-up with no monitoring, he receives an additional informational rent. It turns out that in this case the agent may decide not to screen for the types of the subcontractor, by offering a pooling subcontract he ensures the participation of the bottom agent without requiring any information transmission.

The freedom of the first agent in deciding which type of subcontract to offer is another element of conflicting interest in the model, screening for the type of the second agent is a costly activity and he must be given incentives to perform it. Technically this will introduce a moral-hazard dimension in our model and will reduce the efficiency of the organization despite the monitoring by the principal.

In other words the mark-up on contracting costs is now substituted by the incentives to screen the subcontractor’s type, these costs are actually smaller and therefore overall the principal benefits from the monitoring and there is an efficiency gain for the organization with respect to the non-monitoring case although all these gains

are lower than those we would observe if the principal could force the contractor to choose a particular form of subcontract.

This work is in the stream of literature on collusion and delegation in hierarchies which started with Tirole [1986]¹ that gave a clear cut to the way in which organizations and hierarchies were studied in economic theory. They were no longer considered single blocks but networks of overlapping and nested principal-agent relationships where coalition formation and side-contracting are allowed. For a recent overview of the thriving literature studying the additional incentives problems that delegation and collusion can cause in very simple hierarchies see Mookherjee [2003].

Our set-up instead comes from an extension of Laffont and Martimort [1998] where they compare decentralized and centralized organization of a production process when there are limits on communication.

An analysis very similar to ours is carried out in Baron and Besanko [1992] but they do not model the possibility of a reaction by the agent through the choice of subcontract. New to our paper is in fact the endogenization of the informational structure in the hierarchy, that depends on the actions chosen by the agent.

Most of the delegation literature has considered monitoring by an unproductive agent who, through a costly or costless audit, learns the type of the productive agent and then reports to the principal (see Tirole [1986] for hard information case and Faure-Grimaud, Laffont and Martimort [2003] for a soft information example).

Dequiedt and Martimort [2004] analyze the case of a productive agent who can acquire soft information. Their setting is a hierarchy where the first productive agent can choose whether to learn the type of the second agent through fixed cost monitoring or via arm's length contracting. The choice affects the overall costs of information acquisition and the distribution of rents in the hierarchy. They then study how the optimal contract, designed by the principal, changes with the cost of monitoring. They also have an element of moral hazard in the model because the preferences over the information acquisition methods of the principal and the agent may not be aligned.

The structure of the paper is as follows. Section 2 presents the model, utility functions and contracts. Section 3 derives the optimal delegation proof contract in the benchmark case. Section 4 studies the same organizational structure but allows for the monitoring by the principal. Section 6 concludes. All proofs are in Appendix.

2 The Model.

The principal P wants to buy a quantity q of final output. The two agents, A_i ($i = 1, 2$), produce inputs q_i ($i = 1, 2$) which are needed to produce the final good.

¹On collusion in hierarchies see also Tirole [1992] and Laffont and Martimort [1997, 2000].

These inputs are perfect complements so that $q = q_1 = q_2^2$.

Each agent A_i ($i = 1, 2$) faces a constant marginal cost θ_i of producing good i . These marginal costs are independently drawn from the same common knowledge distribution with discrete support $\Theta_i = \Theta = \{\underline{\theta}, \bar{\theta}\}$, and $\Delta\theta = \bar{\theta} - \underline{\theta} > 0$. With probability ν the agent is efficient, i.e. $\theta_i = \underline{\theta}$. With probability $(1 - \nu)$ the agent is inefficient, i.e. $\theta_i = \bar{\theta}$.

Each agent knows only its own cost and not that of the other agent, while the principal is uninformed on both agents' costs.

The principal maximizes her revenue minus the monetary transfer to the first agent:

$$W = S(q) - t$$

where $S'(\cdot) > 0$, $S''(\cdot) < 0$, satisfies Inada conditions ($S'(0) = +\infty$, $S'(+\infty) = 0$) and $S(0) = 0$.

The principal contracts directly with the prime contractor A_1 and delegates to him the task of contracting with the subcontractor A_2 .

The first agent's utility is given by the monetary transfer received by the principal minus the total costs:

$$U_1 = t - \theta_1 q - y$$

where y is the transfer he makes to the second agent at the subcontracting stage. The second agent's utility is given by:

$$U_2 = y - \theta_2 q$$

If we had a centralized structure (where the principal directly contracts with each agent) we would obtain the following second best³ quantities and rents:

- $S'(q(\underline{\theta}, \underline{\theta})) = 2\underline{\theta}$
- $S'(q(\underline{\theta}, \bar{\theta})) = S'(q(\bar{\theta}, \underline{\theta})) = S'(\hat{q}) = \underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta$
- $S'(q(\bar{\theta}, \bar{\theta})) = 2\bar{\theta} + \frac{2\nu}{1-\nu}\Delta\theta$
- $U_1(\bar{\theta}, \theta_i) = U_2(\theta_i, \bar{\theta}) = 0$
- $U_1(\underline{\theta}, \theta_i) = U_2(\theta_i, \underline{\theta}) = \Delta\theta(\nu q(\underline{\theta}, \bar{\theta}) + (1 - \nu)q(\bar{\theta}, \bar{\theta}))$

In a centralized organization agents are treated symmetrically by the principal and obtain a positive informational rent only when they are efficient.

²In other words the production process is componetised. As in Baron and Besanko [1992] we use the word *componetised* in the sense that the good is formed by putting together components in fixed proportions. The components are produced by different firms or organizational units. As an example we can think of a producer of electricity and a distributor of electricity.

³Laffont and Martimort [1997] show that this outcome is also collusion proof.

2.1 The contracts

As we mentioned in the previous section the organization of the productive activity is decentralized, the principal contracts with A_1 and then the latter contracts with A_2 . Therefore we will have to study two contracts, which will be offered by the parties at different stages.

The principal proposes a grand contract, GC , to the first agent that specifies a quantity to be produced and a transfer, i.e. a pair $\left\{q\left(\widehat{\theta}_1, \widehat{\theta}_2\right), t\left(\widehat{\theta}_1, \widehat{\theta}_2\right)\right\}$, where $q(\cdot)$ is total output, $t(\cdot)$ is the transfer from P to A_1 and $\left(\widehat{\theta}_1, \widehat{\theta}_2\right)$ the report made by A_1 after he has subcontracted with A_2 .

At a later stage, A_1 , who is the one allowed to communicate with A_2 , offers a subcontract, SC , to the second agent that consists of a *manipulation-function*⁴ of reports and a transfer, i.e. $\left\{\Phi\left(\theta_1, \widetilde{\theta}_2\right), y\left(\theta_1, \widetilde{\theta}_2\right)\right\}$, where $\widetilde{\theta}_2$ is the report from A_2 to A_1 . The subcontract thus allows the agents to coordinate the reports to P , reallocate payments and possibly production assignments between themselves.

Throughout the paper we assume that subcontracting is not contractible, that is the contract between the principal and the first agent cannot specify a particular subcontract between the two agents.

In order to simplify notation, denote $t(\bar{\theta}, \bar{\theta}) = \bar{t}$; $t(\underline{\theta}, \bar{\theta}) = \widehat{t}_1$; $t(\bar{\theta}, \underline{\theta}) = \widehat{t}_2$; $t(\underline{\theta}, \underline{\theta}) = \underline{t}$ and use a similar notation for $q(\cdot)$.

2.2 The timing.

The timing of the game is the following:

1. Nature draws θ_i each agent learns his cost.
2. P proposes the grand contract M to A_1 .
3. A_1 offers SC to A_2 .
4. A_2 accepts or refuses the other agent's offer, if he refuses the game ends and both agents get their reservation utility.
5. A_2 reports to A_1 .
6. A_1 accepts or refuses M , if he refuses the game ends.
7. A_1 reports to P according to the manipulation function $\Phi\left(\theta_1, \widetilde{\theta}_2\right)$.
8. Output and monetary transfers are implemented. t to A_1 according to M . y to A_2 according to SC .

⁴This is a function that to any true pair of types assigns a pair of messages to be delivered to the principal $\Phi: \Theta^2 \rightarrow M_1 \times M_2$ (we do not allow random messages). Then because of the Revelation Principle the relevant range for $\Phi(\theta_1, \theta_2)$ will be Θ^2 .

The play of the game is such that the first agent decides on participation only after receiving the report from the second agent. In other words, he knows the state of the world and his individual rationality constraints will be ex-post, resulting in higher costs for the principal. Ex-post participation has the same effect of assuming limited liability or risk aversion⁵.

Alternatively we could have modeled participation decision by A_1 before the contracting with A_2 , in which case delegation would have been equivalent to centralization⁶.

In our setting instead, A_1 has a double advantage over the principal at the acceptance stage. He knows two pieces of information and to report them truthfully he will require more than twice the “standard” informational rent. The choice of this timing is consistent with our intention of dealing with an environment that is not equivalent to a centralized structure and where delegation is truly costly.

Moreover if the principal leaves the middle agent in charge of contracting with his supplier it is unlikely that she will be able to prevent them from communicating before accepting the grand contract. This timing is particularly relevant for short-term projects that do not commit suppliers for a very long period of time. It is quite plausible that before accepting to enter into a new venture the contractor will want to contract with the subcontractor.

3 Benchmark model of delegation.

In this section we analyze the contracts that constitute an equilibrium in a simple framework of hierarchical contracting which we will use as benchmark when we introduce monitoring in the next section⁷.

3.1 The side contract.

The game has two stages so we can solve it backwards by starting at the subcontracting stage. When agent A_1 , being of type θ_1 , offers the subcontract to the bottom agent he maximizes his expected utility with respect to a manipulation function and a transfer to the other agent. Contracting takes place under asymmetric information, so participation and incentive compatibility constraints for A_2 have to be considered when solving the following problem, $SC(\theta_1)$:

⁵See for example Faure-Grimaud, Laffont and Martimort [2003] and Faure-Grimaud and Martimort [2001].

⁶This is a well established result (see for example Laffont and Martimort [1998]). If the agent accepts the contract without knowing the type of the other agent then individual rationality constraint have to be satisfied at interim. There is no asymmetric information between P and A_1 regarding the type of A_2 , hence, given risk neutrality of agents, the reports of the two types will be obtained at no additional cost compared to centralisation.

⁷The analysis of this section follows an extension of Laffont and Martimort [1998].

$$SC(\theta_1) = \begin{cases} \max_{\substack{\Phi(\theta_1, \theta_i) \\ y(\theta_1, \theta_i)}} E_{\theta_2} [U_1(\theta_1)] = \nu (t(\Phi(\theta_1, \underline{\theta})) - y(\theta_1, \underline{\theta}) - \theta_1 q(\Phi(\theta_1, \underline{\theta}))) \\ \quad + (1 - \nu) (t(\Phi(\theta_1, \bar{\theta})) - y(\theta_1, \bar{\theta}) - \theta_1 q(\Phi(\theta_1, \bar{\theta}))) \\ \text{s.t.} \\ y(\theta_1, \bar{\theta}) - \bar{\theta} q(\Phi(\theta_1, \bar{\theta})) = 0 \\ y(\theta_1, \underline{\theta}) - \underline{\theta} q(\Phi(\theta_1, \underline{\theta})) = y(\theta_1, \bar{\theta}) - \underline{\theta} q(\Phi(\theta_1, \bar{\theta})) \end{cases} \quad (1)$$

The above two constraints are the participation constraint of an inefficient second agent and the incentive compatibility constraint of an efficient one respectively, the other constraints are satisfied if the schedule of output is monotone. They are ex-post constraints because the subcontractor perfectly knows the state of the world since the offer by the contractor is revealing of his own type⁸. Rearranging the two binding constraints we obtain the transfers to the bottom agent:

$$y(\theta_1, \bar{\theta}) = \bar{\theta} q(\Phi(\theta_1, \bar{\theta})) \quad (2)$$

$$y(\theta_1, \underline{\theta}) = \underline{\theta} q(\Phi(\theta_1, \underline{\theta})) + \Delta\theta q(\Phi(\theta_1, \bar{\theta})) \quad (3)$$

These transfers are conditional on the reported type and the joint report to the principal and leave some rent to the efficient subcontractor.

3.2 The Grand Contract.

When offering the grand contract the principal is presented with a more complicated problem than when she deals with just one agent who does not interact with other players of the game. The first agent has a double informational advantage at the time of reporting and the principal wants him to reveal truthfully both types.

Incentive compatibility constraints for A_1 are quite resemblant to the coalition incentive compatibility ones of the collusion literature because they take into account the rents paid from one agent to the other at the subcontracting stage. We are going to apply the *Delegation-Proofness Principle*⁹, a variant of the Revelation

⁸Since the first agent has private information and acts as a principal when contracting with the bottom agent we are in an informed principal framework. As Maskin and Tirole [1990] have shown when utility functions are quasilinear the principal cannot gain from concealing her private information. Therefore A_1 does not lose from making a revealing offer, i.e. offer a sub-contract which is dependent on his type.

⁹As is becoming common in the works on delegation we loosely borrow from the collusion literature and the concept of collusion proofness, for a definition see Tirole [1992]. In the collusion framework the null side-contract involves also no transfers between the agent, this of course cannot happen in delegation models where transfers are legitimate. For definition and application of Delegation Proofness and its link with Collusion Proofness see Laffont and Martimort [1998] and Faure-Grimaud, Laffont and Martimort [2003].

Principle, that states that there is no loss of generality in restricting attention to the study of contracts which are unchanged through the process of delegation, i.e. such that the optimal subcontract is the “null subcontract” that is a contract where the manipulation function is equal to the identity function ($\Phi(\theta_1, \tilde{\theta}_2) = (\theta_1, \tilde{\theta}_2)$).

The following Lemma states the conditions under which a grand contract is delegation proof in our framework.

Lemma 1 *A grand contract, GC, is weakly delegation proof if the following incentive compatibility constraints are satisfied:*

$$t(\underline{\theta}, \underline{\theta}) - 2\underline{\theta}q(\underline{\theta}, \underline{\theta}) \geq t(\theta_1, \theta_2) - 2\underline{\theta}q(\theta_1, \theta_2) \quad (4)$$

$$t(\bar{\theta}, \underline{\theta}) - (\bar{\theta} + \underline{\theta})q(\bar{\theta}, \underline{\theta}) \geq t(\theta_1, \theta_2) - (\bar{\theta} + \underline{\theta})q(\theta_1, \theta_2) \quad (5)$$

$$t(\underline{\theta}, \bar{\theta}) - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)q(\underline{\theta}, \bar{\theta}) \geq t(\theta_1, \theta_2) - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)q(\theta_1, \theta_2) \quad (6)$$

$$t(\bar{\theta}, \bar{\theta}) - \left(2\bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)q(\bar{\theta}, \bar{\theta}) \geq t(\theta_1, \theta_2) - \left(2\bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)q(\theta_1, \theta_2) \quad (7)$$

$\forall (\theta_1, \theta_2) \in \Theta \times \Theta$.

The above constraints give the conditions that the transfers to the middle agent have to satisfy to obtain truthful report in the grand contract in each possible state of nature. The reports are going to be ex-post efficient only for pairs that involve an efficient second agent, in the other two cases there is some inefficiency due to the asymmetric information at the subcontract stage. In particular a coalition of the kind $(\bar{\theta}, \underline{\theta})$ is more efficient from the point of view of the principal than a coalition of the kind $(\underline{\theta}, \bar{\theta})$ ¹⁰, and the former has an incentive to mimic the latter. This difference is due to the fact that A_1 has paid informational rent to an efficient second agent, ultimately we observe a reduction in efficiency due to the cascading of informational rents down the layers of the hierarchy¹¹. The additional mark-up on subcontracting costs that the prime contractor is able to get is exactly the source of the costs of delegation when we compare it to a centralized setting.

When choosing the optimal contract the principal will maximize her expected utility over the four possible contractor-subcontractor pairs, that is:

¹⁰The virtual type of a coalition $(\bar{\theta}, \underline{\theta})$ is lower than the virtual type of a coalition $(\underline{\theta}, \bar{\theta})$, where the virtual type is the relevant type for the principal when she chooses production assignments and it is given by the actual type plus the informational rent.

¹¹The informational rent for an efficient second agent is $\Delta\theta q(\Phi(\theta_1, \bar{\theta}))$ so distorting upward the report will reduce the quantity prescribed for a pair $(\theta_1, \bar{\theta})$, but also cause a decrease of the informational rent paid by A_1 in the other two possible situations $(\theta_1, \underline{\theta})$.

$$\begin{aligned} \max_{E_{\theta_1, \theta_2}} [W] &= \nu^2 (S(\underline{q}) - \underline{t}) + \nu(1-\nu) (S(\widehat{q}_1) - \widehat{t}_1) + & (8) \\ &+ \nu(1-\nu) (S(\widehat{q}_2) - \widehat{t}_2) + (1-\nu)^2 (S(\bar{q}) - \bar{t}) \end{aligned}$$

Subject to incentive compatibility constraints (4-7) and the following ex-post individual rationality constraints:

$$\begin{aligned} \underline{t} - 2\underline{\theta}\underline{q} - \Delta\theta\widehat{q}_1 &\geq 0 \\ \widehat{t}_1 - (\underline{\theta} + \bar{\theta})\widehat{q}_1 &\geq 0 \\ \widehat{t}_2 - (\underline{\theta} + \bar{\theta})\widehat{q}_2 - \Delta\theta\bar{q} &\geq 0 \\ \bar{t} - 2\bar{\theta}\bar{q} &\geq 0 \end{aligned}$$

Conditional on the optimal schedule of output being monotone we can restrict attention to the following binding constraints, knowing that the others will be also satisfied:

$$\begin{aligned} \underline{t} - 2\underline{\theta}\underline{q} &= \widehat{t}_2 - 2\bar{\theta}\widehat{q}_2 \\ \widehat{t}_2 - (\underline{\theta} + \bar{\theta})\widehat{q}_2 &= \widehat{t}_1 - (\underline{\theta} + \bar{\theta})\widehat{q}_1 \\ \widehat{t}_1 - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\widehat{q}_1 &= \bar{t} - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\bar{q} \\ \bar{t} - 2\bar{\theta}\bar{q} &= 0 \end{aligned}$$

These considerations simplify the optimization problem considerably, from the constraints we obtain the incentive feasible transfers which allow to solve for the optimal contract, as stated in the following proposition.

Proposition 1 *The optimal delegation proof contract has the following properties:*

- for $\nu < \nu^*$
 - It implements a decreasing schedule of outputs $\underline{q} > \widehat{q}_2 > \widehat{q}_1 > \bar{q}$ where the prescribed quantities are implicitly defined by:
 - * $S'(\underline{q}) = 2\underline{\theta}$
 - * $S'(\widehat{q}_2) = \bar{\theta} + \underline{\theta} + \frac{\nu}{1-\nu}\Delta\theta$
 - * $S'(\widehat{q}_1) = \bar{\theta} + \underline{\theta} + \frac{\nu(2-\nu)}{(1-\nu)^2}\Delta\theta$
 - * $S'(\bar{q}) = 2\bar{\theta} + \frac{\nu(2-\nu)(1-2\nu)}{(1-\nu)^3}\Delta\theta$
 - The informational rents granted to the agents are the following:
 - * $U_1(\underline{\theta}, \underline{\theta}) = \Delta\theta(\widehat{q}_2 - \widehat{q}_1) + \frac{\nu}{1-\nu}\Delta\theta\widehat{q}_1 + \frac{1-2\nu}{(1-\nu)}\Delta\theta\bar{q}$
 - * $U_1(\underline{\theta}, \bar{\theta}) = \Delta\theta\bar{q} + \frac{\nu}{1-\nu}\Delta\theta(\widehat{q}_1 - \bar{q})$

$$\begin{aligned}
& * U_1(\bar{\theta}, \underline{\theta}) = \frac{\nu}{1-\nu} \Delta\theta (\hat{q}_1 - \bar{q}) \\
& * U_1(\bar{\theta}, \bar{\theta}) = 0 \\
& * U_2(\underline{\theta}, \underline{\theta}) = \Delta\theta \hat{q}_1 \\
& * U_2(\bar{\theta}, \underline{\theta}) = \Delta\theta \bar{q} \\
& * U_2(\theta_i, \bar{\theta}) = 0
\end{aligned}$$

- for $\nu \geq \nu^*$

– It implements a decreasing schedule of outputs with some bunching $\underline{q} > \hat{q}_2 > \tilde{q} = \hat{q}_1 = \bar{q}$ where the prescribed quantities are implicitly defined by:

$$\begin{aligned}
& * S'(\underline{q}) = 2\underline{\theta} \\
& * S'(\hat{q}_2) = \bar{\theta} + \underline{\theta} + \frac{\nu}{1-\nu} \Delta\theta \\
& * S'(\tilde{q}) = 2\bar{\theta} + \frac{\nu}{(1-\nu)} \Delta\theta
\end{aligned}$$

– The informational rents granted to the agents are the following:

$$\begin{aligned}
& * U_1(\underline{\theta}, \underline{\theta}) = \Delta\theta \hat{q}_2 \\
& * U_1(\underline{\theta}, \bar{\theta}) = \Delta\theta \tilde{q} \\
& * U_1(\bar{\theta}, \underline{\theta}) = U_1(\bar{\theta}, \bar{\theta}) = 0 \\
& * U_2(\underline{\theta}, \underline{\theta}) = U_2(\bar{\theta}, \underline{\theta}) = \Delta\theta \tilde{q} \\
& * U_2(\theta_i, \bar{\theta}) = 0
\end{aligned}$$

This contract requires quantities that are more downward distorted than those in the second best, the amount of informational rent is more than double and consequently the principal optimally trades off some productive efficiency. Comparing these quantities to the second best schedule reveals that the further distortions are in the quantities prescribed to pairs where an inefficient second agent is present, this is due to the extra incentive that A_1 must be given to truthfully report the pair of types after he has paid the informational rent to A_2 . This clearly identifies where the cost for the principal of not being able to communicate directly with one agent lies and it highlights precisely what is meant by the cost of delegation. Since the first agent accepts the contract offered by the principal only after he has learned the type of the second agent, he is given double rent plus a “reimbursement” for the rent he has paid to the second agent.

By comparing the equilibrium payoffs instead, we can see that the bottom agent is treated as in the second best contract: he gets positive rents only when he is efficient. It is different what happens to the informational rent of the first agent when $\nu < \nu^*$, in this sub-case he obtains a positive rent also when he is inefficient but he is paired with an efficient second agent, this is due to the ex-post acceptance of the grand contract that gives him a double informative advantage when deciding about participation in the grand contract. When $\nu \geq \nu^*$ the probability of facing

an efficient agent increases therefore the principal gains by bunching the contracts which involve an inefficient second agent because the screening of a coalition of the $(\underline{\theta}, \bar{\theta})$ type proves too costly.

4 Delegation with Monitoring.

We now assume that the principal can monitor the communication between the contractor and the subcontractor. In other words she observes what goes on at stages 3, 4 and 5 of the game: subcontract offer, subcontract acceptance and report of information by the subcontractor.

It is as if the principal sent a person of trust to be present but silent at the subcontracting stage, we now have a mismatch between the organizational and the informational structure. This is because by monitoring the communication between the agents the principal will potentially learn a lot of private information.

First of all if the subcontract offer is revealing the principal learns the type of the contractor, and, as a consequence, will not offer any rent to him to report to her. This will leave an efficient first agent with a payoff that does not exceed his reservation utility.

In addition, by observing the report that the subcontractor makes to the contractor, the principal will learn the type of the second agent at the same time as the first agent. This implies that the principal is not willing to pay the extra rent to the middle agent to truthfully reveal the type of the bottom agent, therefore saving on what we called the “true” cost of delegation.

It is evident that the subcontractor is not affected by the monitoring activity, he obtains his rents through the subcontract offered by the contractor and does not deal directly with the monitoring principal. It is the first agent who is damaged the most by the monitoring, he could be left with no rent at all in any state of the world.

What is likely to happen then is that at stage 3, when the contractor moves and offers the subcontract he will change his action in an attempt to conceal some information and get some informational rent back.

The first and most obvious reaction would be to conceal his own type when offering the subcontract, this comes at no cost to him and would restore asymmetric information, at least partially, between the first agent and the principal at stage 7 when he reports into the grand-contract.

The way to implement a not revealing contract offer is to offer two menus of two transfer-quantity pairs, each one of them designed for either an efficient or inefficient subcontractor. As Maskin and Tirole [1990] show there is no advantage in doing so vis-a-vis the subcontractor, but in this particular application the gain comes from the relationship with the upper layer of the hierarchy who in spite of observing the subcontract offered does not learn anything about the contractor’s type.

More precisely, while the bottom agent's constraints will be in expected terms (he does not know the middle agent's type) the transfers offered will be the same as in the benchmark case, the ones that satisfy ex-post constraints.

In other words the transfers included in a not-revealing subcontract offer will be:

$$SC(\theta_1, \underline{\theta}) = \{y(\theta_1, \underline{\theta}), \Phi(\theta_1, \underline{\theta}); \theta_1 \in \Theta\}$$

$$SC(\theta_1, \bar{\theta}) = \{y(\theta_1, \bar{\theta}), \Phi(\theta_1, \bar{\theta}); \theta_1 \in \Theta\}$$

These two contracts are designed for an efficient and an inefficient second agent (respectively) but are conditioned on the type of the first agent as well. Any type of the second agent will choose the contract designed for himself and wait until stage 8 to find out exactly what price-quantity pair of the possible two will be implemented.

Note that the subcontractor is as well off as with a revealing subcontract offer so he will not object in any way to this new offer by the contractor.

This is nothing more than an application of Myerson's well known "inscrutability principle" (Myerson [1983]), in fact in this set-up the gain to the offer-maker does not come from the recipient of the offer but at the expenses of the monitor of the offer.

At this stage the monitor no longer learns the type of the contractor but nonetheless still can observe the report done by the subcontractor regarding his type. This means that at the next stage when the contractor reports into the grand-contract his freedom is much limited, he can't misreport the type of the subcontractor which is now common knowledge. The principal therefore saves on the additional rent that had to be given to the contractor to report two pieces of information.

This will be reflected in the grand-contract offer, now P has to give incentives to A_1 to reveal only one piece of information, his own type, because she already knows the type of the second agent. Since the agent cannot misreport the other's type, incentive compatibility needs to hold over two separate pairs of contracts, each pair pools across the types of the second agent.

Lemma 2 *When the principal can monitor the report from A_2 to A_1 , a grand contract is incentive compatible (delegation proof) if the output schedule is monotonic and the following constraints are satisfied:*

$$\underline{t} - 2\underline{\theta}\underline{q} \geq \hat{t}_2 - 2\underline{\theta}\hat{q}_2 \tag{9}$$

$$\hat{t}_1 - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu} \Delta\theta \right) \hat{q}_1 \geq \bar{t} - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu} \Delta\theta \right) \bar{q} \tag{10}$$

Those above are the two constraints which refer to pairs where the first agent is efficient, these are in fact the only relevant ones since there is common knowledge about the type of the second agent at the time of reporting by A_1 . The first agent

knows it because of the side-contract stage and the report it entails, the principal in turn is allowed to listen (or observes) to the truthful report that A_2 makes to A_1 .

The principal must still ensure the participation of the first agent into the grand contract and the set of constraints that have to be satisfied is not different from the benchmark case, only now two of them will be binding while before only one was. The same type of considerations that lead us to the reduction of the number of incentive compatibility constraints are at work here. Now, that the principal monitors and gets to know the type of A_2 , an inefficient first agent will be left with his reservation utility irrespectively of the type of second agent he is matched with. The binding constraints are:

$$\widehat{t}_2 - (\underline{\theta} + \bar{\theta}) \widehat{q}_2 - \Delta\theta\bar{q} = 0 \quad (11)$$

$$\bar{t} - 2\bar{\theta}\bar{q} = 0 \quad (12)$$

In other words the principal is extracting only one piece of information, she knows the type of A_2 and she is not giving any *extra* rent to A_1 to reveal that the second agent is efficient.

It is evident that the monitoring procures benefits to the principal if the contractor is willing to screen for the types of the subcontractor and receives a report about his private information. If the principal could force the middle agent to offer a particular type of subcontract then it would be a screening one and the following proposition summarizes the results in that case.

Proposition 2 *When the principal can costlessly and perfectly monitor the report of the second agent into the subcontract and can force A_1 to offer a screening subcontract the optimal grand contract has the following characteristics:*

- *It implements a decreasing schedule of output $\underline{q} > \widehat{q} > \bar{q}$ (where $\widehat{q} = \widehat{q}_1 = \widehat{q}_2$) implicitly defined by:*

$$\begin{aligned} - S'(\underline{q}) &= 2\underline{\theta} \\ - S'(\widehat{q}) &= (\underline{\theta} + \bar{\theta}) + \frac{\nu}{1-\nu}\Delta\theta \\ - S'(\bar{q}) &= 2\bar{\theta} + \frac{2\nu}{1-\nu}\Delta\theta \end{aligned}$$

- *The informational rents granted to the agents are the following:*

$$\begin{aligned} - U_1(\underline{\theta}, \underline{\theta}) &= \Delta\theta\bar{q} \\ - U_1(\underline{\theta}, \bar{\theta}) &= \frac{\nu}{1-\nu}\Delta\theta\widehat{q} + \frac{1-2\nu}{(1-\nu)}\Delta\theta\bar{q} \\ - U_1(\bar{\theta}, \theta_i) &= 0 \\ - U_2(\underline{\theta}, \underline{\theta}) &= \Delta\theta\widehat{q} \\ - U_2(\bar{\theta}, \underline{\theta}) &= \Delta\theta\bar{q} \end{aligned}$$

$$- U_2(\theta_i, \bar{\theta}) = 0$$

In each state of the world the quantities produced are equal to those that would be produced in a centralized organization, this means that if the principal is allowed to monitor the report made into the subcontract the second best can be achieved.¹² The principal though, cannot do better than the second best even if she gets to know some private information for free, because she receives this information when the second agent is reporting to the first one after he has been given the necessary incentives to do so. These in turn are costs for A_1 that the principal has to reimburse if she wants to ensure the participation of A_1 (and indirectly of A_2 as well) in the production process. In other words, in the overall organization two pieces of private information are to be reported truthfully, exactly the same number as in a centralized setting where both pieces are extracted by the principal.

With the monitoring the extra-cost of delegation compared to centralization disappears, but nothing more: even if informational delegation no longer exists, the principal still faces two agents that have private information and this keeps the model in a second best world.

Taking a closer look at the equilibrium payoffs, it emerges that the expected rents that the principal has to pay are exactly the second best ones, what is different is their distribution: when the principal monitors, an efficient first agent gets more rent when paired with an inefficient second one and less when the other agent is efficient if compared to the payoff in a centralized organization.

If we look at the rents of all the players we can see that the principal gains from the monitoring while the prime contractor is worse off, he receives lower informational rents in two states of the world. The subcontractor at the bottom of the hierarchy, instead, is unaffected by the monitoring because he still receives the right incentives to reveal his type when he subcontracts with the middle agent. It is only after the subcontract has been offered and accepted that the principal gets to know his type.

Unfortunately for the principal the type of subcontract cannot be (by assumption) contracted upon and as a consequence she cannot force the middle agent to offer a screening subcontract to the bottom one.

The aim of the contractor is to eliminate the communication that is being monitored by the principal and restore some freedom when reporting into the grand-contract¹³. In other words A_1 could offer a contract that does not require a report and that is independent from A_2 's type; we call this a pooling subcontract, because it does not separate the types of A_2 . Since the agent might in fact prefer the offer

¹²Note in fact that $\hat{q}_1 = \hat{q}_2$, symmetry is back in the model because the principal can avoid paying the extra-rent so that the two pairs $(\underline{\theta}, \bar{\theta})$ and $(\bar{\theta}, \underline{\theta})$ can now be treated equally as in a centralized organization.

¹³A secret code or an encrypting technology would not help in this situation, the key to this language would need to be included in the subcontract which is monitored by the principal.

of a pooling subcontract he needs to be given the incentive to pick a screening subcontract. In a pooling subcontract the first agent will offer a set of transfers to the second one as if he was always inefficient, namely:

$$y(\theta_1, \theta_2) = \bar{\theta}q(\Phi(\theta_1, \theta_2)).$$

The idea is that by paying always the high marginal cost of production he ensures that both types of A_2 are willing to participate since their individual rationality constraints are satisfied:

$$\bar{U}_2 = 0$$

$$\underline{U}_2 = \Delta\theta q(\Phi(\theta_1, \theta_2)) > 0$$

Because A_2 can only be of two types these ex-post payoffs are the same as in the previous cases, when the incentive compatibility of the efficient type was binding making him indifferent between telling the truth and claiming to be inefficient.

To keep the notation homogeneous we still write $y(\cdot)$ and the quantities to be produced as if they were dependent on both types. Actually in this case the message space for the first agent when reporting to the principal is larger than before (it is equal to the one in the benchmark case), when, because of the monitoring, A_1 was restricted to the message space $\{\theta_i, \theta_2\}$ (he had to report the true θ_2). Now the message space is in fact $\{\theta_i, \theta_j\}$ (with $i, j = 1, 2$) but the pooling contract implies that $\theta_j = \bar{\theta}$ always. As a consequence of this pooling contract the manipulation function is reduced once again to a function of one variable: $\Phi(\theta_1, \theta_2) = (\hat{\theta}_1, \bar{\theta})$.

We need to understand which subcontract offer will be optimal for A_1 . Solving backward the analysis is as follows: for a given grand-contract the agent decides whether to screen or not, then the principal optimally sets the terms of the grand-contract.

Given a grand-contract $GC = \{t, q, \hat{t}_1, \hat{q}_1, \hat{t}_2, \hat{q}_2, \bar{t}, \bar{q}\}$ the contractor will choose the type of subcontract that maximizes his expected utility.

The expected utility when offering a separating subcontract is:

$$\begin{aligned} U_1(\theta_1) = & \nu(t(\Phi(\theta_1, \underline{\theta})) - \underline{\theta}q(\Phi(\theta_1, \underline{\theta})) - \Delta\theta q(\Phi(\theta_1, \bar{\theta})) - \theta_1 q(\Phi(\theta_1, \underline{\theta}))) + \\ & (1 - \nu)(t(\Phi(\theta_1, \bar{\theta})) - \bar{\theta}q(\Phi(\theta_1, \bar{\theta})) - \theta_1 q(\Phi(\theta_1, \bar{\theta}))) \end{aligned}$$

while the expected utility when offering a pooling subcontract is:

$$U_P(\theta_1) = t(\Phi(\theta_1, \bar{\theta})) - (\bar{\theta} + \theta_1)q(\Phi(\theta_1, \bar{\theta})) .$$

The principal will require truthful revelation and a separating subcontract.

The incentive compatibility constraints will be of the form seen previously in this section, because of the monitoring the principal will know the type of the second

agent (if the subcontract requires a report) and the incentives will be for the first agent to report only his own type.

In order for the subcontract offered to be a screening one a new constraint will have to be satisfied, the expected payoff from such a contract offer will have to be higher than the one secured by a pooling subcontract, more precisely:

$$\nu U_1(\underline{\theta}, \underline{\theta}) + (1 - \nu) U_1(\underline{\theta}, \bar{\theta}) \geq U_P^*(\underline{\theta}_1) \quad (13)$$

where $U_1(\underline{\theta}, \underline{\theta})$ and $U_1(\underline{\theta}, \bar{\theta})$ are the rents earned by an efficient first agent who is paired with an efficient and inefficient second agent respectively when he offers a separating subcontract and truthfully reports to the principal¹⁴. While $U_P^*(\underline{\theta}_1)$ is the maximum utility that can be achieved by an efficient first agent that offers a pooling subcontract, and it is defined as:

$$U_P^*(\underline{\theta}_1) = \max_{\Phi} (\Phi(\underline{\theta}_1, \bar{\theta})) - (\bar{\theta} + \underline{\theta}_1) q(\Phi(\underline{\theta}_1, \bar{\theta}))$$

Constraint (13) is in fact a moral hazard constraint, when designing the contract the principal has to give incentives to the first agent to do her preferred action which, in this case, is offering a screening contract. Since communication is observed by the principal screening the bottom agent becomes a costly activity of which the middle agent does not reap all the benefits, he must be given incentives to perform it.

It is worth noting that $U_P^*(\underline{\theta}_1)$ could be achieved by truthtelling but also by any other report, the following Remark is of some help in this direction.

Remark 1 *If a Grand Contract is incentive compatible when the subcontract offer is separating then it is incentive compatible if the offer is pooling and the expected utility of A_1 is:*

$$U_P^*(\underline{\theta}_1) = \hat{t}_1 - (\underline{\theta} + \bar{\theta}) \hat{q}_1.$$

Having calculated the maximum the contractor can obtain with any of the two possible contract offer we can now summarize the set of constraints that the grand-contract will have to satisfy to be delegation-proof and to induce a separating subcontract offer.

Lemma 3 *When the principal can monitor the report from A_2 to A_1 , a grand contract is incentive compatible (delegation proof) and will induce a separating subcon-*

¹⁴We can, without loss of generality, limit the analysis to the case of an efficient first agent because an inefficient one will receive his reservation utility regardless of the type of sub-contract offered.

tract if the output schedule is monotonic and the following constraints are satisfied:

$$\begin{aligned}
\bar{t} - 2\bar{\theta}\bar{q} &\geq 0 \\
\hat{t}_2 - (\underline{\theta} + \bar{\theta})\hat{q}_2 - \Delta\theta\bar{q} &\geq 0 \\
\underline{t} - 2\underline{\theta}\underline{q} &\geq \hat{t}_2 - 2\theta\hat{q}_2 \\
\hat{t}_1 - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\hat{q}_1 &\geq \bar{t} - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\bar{q} \\
\underline{t} - 2\underline{\theta}\underline{q} &\geq \hat{t}_1 - 2\theta\hat{q}_1
\end{aligned}$$

The first two constraints are the participation constraints of the two coalitions in which an inefficient contractor is present, because of the monitoring also the mixed coalition in which the subcontractor is efficient is kept at the reservation utility level¹⁵. The next two are the coalition incentive constraints of pairs in which an efficient contractor is present. The last one is the moral-hazard constraint which guarantees that the subcontract offer is separating.

The moral-hazard constraint has in fact the appearance of another incentive constraint that makes an efficient contractor prefer the allocation that he obtains when paired with an efficient subcontractor, this will make him offer a screening subcontract (since that is the only way of having an efficient A_2 into the contract).

The next proposition characterizes the optimal contract when the principal is monitoring and the agent is free to choose the type of subcontract.

Proposition 3 *When the principal can costlessly and perfectly monitor the report of the second agent into the subcontract and can force A_1 to offer a screening subcontract the optimal grand contract has the following characteristics:*

- It implements a decreasing schedule of output $\underline{q} > \hat{q}_2 > \hat{q}_1 > \bar{q}$ (where $\hat{q}_1 = (1-\nu)\hat{q}_2 + \nu\bar{q}$) implicitly defined by:

$$\begin{aligned}
- S'(\underline{q}) &= 2\underline{\theta} \\
- S'(\hat{q}_2) &= (\underline{\theta} + \bar{\theta}) + \frac{\nu}{1-\nu}\Delta\theta \\
- S'(\bar{q}) &= 2\bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta + \frac{\nu(1-\nu)}{1-\nu+\nu^2}\Delta\theta
\end{aligned}$$

- The informational rents granted to the agents are the following:

$$\begin{aligned}
- U_1(\underline{\theta}, \theta_i) &= \Delta\theta[\nu\hat{q}_2 + (1-\nu)\bar{q}] \\
- U_1(\bar{\theta}, \theta_i) &= 0 \\
- U_2(\underline{\theta}, \underline{\theta}) &= \Delta\theta\hat{q}_1 = \Delta\theta[(1-\nu)\hat{q}_2 + \nu\bar{q}] \\
- U_2(\bar{\theta}, \underline{\theta}) &= \Delta\theta\bar{q}
\end{aligned}$$

¹⁵ Again, the contractor gets no rent while the subcontractor receives a positive ex-post payoff.

$$- U_2(\theta_i, \bar{\theta}) = 0$$

The optimal contract in the case the agent is free to choose the type of subcontract when the principal is monitoring generates equilibrium which is in between the second best and the benchmark case for both the principal and agent. This is because taking into account the possibility that the agent offers a pooling subcontract amounts to bringing back into the model some discretion over the report of the type of the second agent. If in the benchmark case the contractor had full flexibility about what to report now he has at least the freedom to ask or not for a report.

The quantities are slightly more distorted than in the second best and because of the newly introduced moral hazard constraint there is some form of bunching, \hat{q}_1 is not optimally determined but it is an average of \hat{q}_2 and \bar{q} .

5 Conclusions and Discussion.

Our analysis of a supplier's hierarchy highlights the importance of information transmission in these contracting relationships. The message is that private information is difficult to obtain for free, once the principal saves on some costs of information transmission she is forced to give incentives for information acquisition.

We have shown that the effectiveness of costless monitoring by the principal is greatly reduced when the monitored party is free to choose the type of subcontract.

We believe this contributes to the literature on the functioning of hierarchies both inside firms or in markets. Despite taking the organizational and contracting structure as given, we were able to endogenize the informational structure and show how it is affected by the decision to monitor. The final efficiency loss is caused by the non-alignment of the preferences for information acquisition of the head of the hierarchy and the contractor.

We have made some simplifying assumptions, most of which are not essential for the results we obtain.

First of all the results go through even when we assume some input substitutability (i.e. Cobb-Douglas production function), the difference is that when inputs are not perfect complements the middle agent will do inefficiently little outsourcing to the bottom agent. This is an additional moral-hazard component on top of all the information distortions which we have seen being exacerbated by delegation and that are precisely the focus of the paper.

Secondarily, the two-type setting greatly simplifies the analysis because it limits the possible subcontracts that the middle agent could choose. We conjecture, but have not proven, that the qualitative results would not change if the type space was richer.

Finally an assumption which is not innocuous is the impossibility of the principal

to condition payments on the type of subcontract offer, this would partially eliminate the ability of the first agent to hide some information. The idea that “subcontracting is not contractible” is in some cases quite plausible especially when it is an unobservable action the result of which (the report from the bottom agent) does not provide a perfect signal of it.

References

- [1] Baron, D.P. and D. Besanko (1992) “Information, Control, and Organizational Structure”, *Journal of Economics & Management Strategy*, 1:237-275.
- [2] Dequiedt V. and D. Martimort (2004) “Delegated Monitoring versus Arm’s Length Contracting”, *International Journal of Industrial Organization*, 22:951-981.
- [3] Faure-Grimaud, A., J.J. Laffont and D. Martimort (2003) “Collusion, Delegation and Supervision with Soft Information”, *Review of Economic Studies*, 70:253-280.
- [4] Faure-Grimaud and D. Martimort (2001) “On some Agency Costs of Intermediated Contracting”, *Economic Letters*, 71:75-81.
- [5] Laffont, J.J. and D. Martimort (1997) “Collusion Under Asymmetric Information”, *Econometrica*, 65:875-911.
- [6] Laffont, J.J. and D. Martimort (1998) “Collusion & Delegation”, *Rand Journal of Economics*, 29:280-305.
- [7] Laffont, J.J. and D. Martimort (2000) “Mechanism Design with Collusion and Correlation”, *Econometrica*, 68:238-263.
- [8] Maskin, E. and J. Tirole (1990) “The Principal-Agent relationship with an Informed Principal: the Case of Private Values”, *Econometrica*, 58:379-409.
- [9] Myerson, R.B. (1983) “Mechanism Design by an Informed Principal”, *Econometrica*, 51:1767-1797.
- [10] Melumad, D.M., D. Mookherjee and S. Reichelstein (1995) “Hierarchical Decentralization of Incentives Contract”, *Rand Journal of Economics*, 26:654-672.
- [11] Mookherjee, D. (2003) “Delegation and Contracting Hierarchies: An Overview”, mimeo, Boston University.
- [12] Mookherjee, D. and S. Reichelstein (1992) “Dominant Strategy Implementation of Bayesian Incentive Compatible Rules”, *Journal of Economic Theory*, 56:378-399.

- [13] Mookherjee, D. and M. Tsumagari (2004) “The Organization of Supplier Networks: Effects of Delegation and Intermediation”, *Econometrica*, 72:1179-1220.
- [14] Tirole, J. (1986) “Hierarchies & Bureaucracies: On the Role of collusion in Organizations”, *Journal of Law, Economics & Organizations*, 2:181-214.
- [15] Tirole, J. (1992) “Collusion and the Theory of Organization”, in J.J. Laffont (ed.), *Advances in Economic Theory*, Cambridge University Press, Cambridge.

Appendix.

Proof of Lemma 1.. If we substitute the optimal side transfers (3) and (2) into the first agent's expected utility function in problem (1) we get:

$$\begin{aligned} \max E_{\theta_2} [U_1] &= \nu (t(\Phi(\theta_1, \underline{\theta})) - \underline{\theta}q(\Phi(\theta_1, \underline{\theta})) - \Delta\theta q(\Phi(\theta_1, \bar{\theta})) - \theta_1 q(\Phi(\theta_1, \underline{\theta}))) + \\ &\quad (1 - \nu) (t(\Phi(\theta_1, \bar{\theta})) - \bar{\theta}q(\Phi(\theta_1, \bar{\theta})) - \theta_1 q(\Phi(\theta_1, \bar{\theta}))) \end{aligned}$$

now we can check for incentive compatibility for any of the possible "coalitions" (there are four of them).

Then the condition for the optimality of $\Phi(\underline{\theta}, \underline{\theta}) = (\underline{\theta}, \underline{\theta})$ is the following:

$$\begin{aligned} \nu (t(\underline{\theta}, \underline{\theta}) - 2\underline{\theta}q(\underline{\theta}, \underline{\theta}) - \Delta\theta q(\Phi(\underline{\theta}, \bar{\theta}))) + (1 - \nu) (t(\Phi(\underline{\theta}, \bar{\theta})) - (\bar{\theta} + \underline{\theta})q(\Phi(\underline{\theta}, \bar{\theta}))) &\geq \\ \nu (t(\theta_1, \theta_2) - 2\underline{\theta}q(\theta_1, \theta_2) - \Delta\theta q(\Phi(\underline{\theta}, \bar{\theta}))) + (1 - \nu) (t(\Phi(\underline{\theta}, \bar{\theta})) - (\bar{\theta} + \underline{\theta})q(\Phi(\underline{\theta}, \bar{\theta}))). \end{aligned}$$

For $\Phi(\underline{\theta}, \bar{\theta}) = (\underline{\theta}, \bar{\theta})$ is:

$$\begin{aligned} \nu (t(\Phi(\underline{\theta}, \underline{\theta})) - 2\underline{\theta}q(\Phi(\underline{\theta}, \underline{\theta})) - \Delta\theta q(\underline{\theta}, \bar{\theta})) + (1 - \nu) (t(\underline{\theta}, \bar{\theta}) - (\underline{\theta} + \bar{\theta})q(\underline{\theta}, \bar{\theta})) &\geq \\ \nu (t(\Phi(\underline{\theta}, \underline{\theta})) - 2\underline{\theta}q(\Phi(\underline{\theta}, \underline{\theta})) - \Delta\theta q(\theta_1, \theta_2)) + (1 - \nu) (t(\theta_1, \theta_2) - (\underline{\theta} + \bar{\theta})q(\theta_1, \theta_2)). \end{aligned}$$

For $\Phi(\bar{\theta}, \underline{\theta}) = (\bar{\theta}, \underline{\theta})$ is:

$$\begin{aligned} \nu (t(\bar{\theta}, \underline{\theta}) - (\bar{\theta} + \underline{\theta})q(\bar{\theta}, \underline{\theta}) - \Delta\theta q(\Phi(\bar{\theta}, \bar{\theta}))) + (1 - \nu) (t(\Phi(\bar{\theta}, \bar{\theta})) - 2\bar{\theta}q(\Phi(\bar{\theta}, \bar{\theta}))) &\geq \\ \nu (t(\theta_1, \theta_2) - (\bar{\theta} + \underline{\theta})q(\theta_1, \theta_2) - \Delta\theta q(\Phi(\bar{\theta}, \bar{\theta}))) + (1 - \nu) (t(\Phi(\bar{\theta}, \bar{\theta})) - 2\bar{\theta}q(\Phi(\bar{\theta}, \bar{\theta}))). \end{aligned}$$

For $\Phi(\bar{\theta}, \bar{\theta}) = (\bar{\theta}, \bar{\theta})$ is:

$$\begin{aligned} \nu (t(\Phi(\bar{\theta}, \underline{\theta})) - (\bar{\theta} + \underline{\theta})q(\Phi(\bar{\theta}, \underline{\theta})) - \Delta\theta q(\bar{\theta}, \bar{\theta})) + (1 - \nu) (t(\bar{\theta}, \bar{\theta}) - 2\bar{\theta}q(\bar{\theta}, \bar{\theta})) &\geq \\ \nu (t(\Phi(\bar{\theta}, \underline{\theta})) - (\bar{\theta} + \underline{\theta})q(\Phi(\bar{\theta}, \underline{\theta})) - \Delta\theta q(\theta_1, \theta_2)) + (1 - \nu) (t(\theta_1, \theta_2) - 2\bar{\theta}q(\theta_1, \theta_2)). \end{aligned}$$

Simplifying we obtain constraints (4)-(7). This are the conditions for truthtelling, what any coalition gets in the contract leaves it better off than anything else they could have gotten misreporting their types.

Note that we have used first agent's expected utility because the manipulation function is part of the side-contract that is offered before getting to know the second agent's type, then incentive compatibility constraints are ex-post because by the time A_1 reports to the principal he knows the true type of A_2 . ■

Proof of Proposition 1. Given the binding constraints we can manipulate them and obtain the incentive compatible and individually rational transfers:

$$\begin{aligned} \underline{t} &= 2\underline{\theta}q + \Delta\theta\hat{q}_2 + \frac{\nu}{1-\nu}\Delta\theta\hat{q}_1 + \frac{1-2\nu}{1-\nu}\Delta\theta\bar{q} \\ \hat{t}_2 &= (\underline{\theta} + \bar{\theta})\hat{q}_2 + \frac{\nu}{1-\nu}\Delta\theta\hat{q}_1 + \frac{1-2\nu}{1-\nu}\Delta\theta\bar{q} \\ \hat{t}_1 &= \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\hat{q}_1 + \frac{1-2\nu}{1-\nu}\Delta\theta\bar{q} \\ \bar{t} &= 2\bar{\theta}\bar{q} \end{aligned}$$

We can substitute them in the principal's objective function and then maximize with respect to q , \hat{q}_1 , \hat{q}_2 and \bar{q} , we then obtain the decreasing schedule of output in the first part of Proposition 1.

But we need to ensure that monotonicity is satisfied, and $\hat{q}_1 > \bar{q}$ is true only when:

$$\bar{\theta} + \underline{\theta} + \frac{\nu(2-\nu)}{(1-\nu)^2} \Delta\theta < 2\bar{\theta} + \frac{\nu(2-\nu)(1-2\nu)}{(1-\nu)^3} \Delta\theta.$$

The above is satisfied when $\nu < \nu^*$ where ν^* is a root of:

$$(1-\nu)^3 - \nu^2(2-\nu) = 0$$

which is $\nu^* = \frac{3}{2} - \frac{1}{2}\sqrt{5} \simeq .38197$.

If $\nu \geq \nu^*$ the the optimal contract requires some pooling. This means that two different pairs will be offered the same contract $\hat{t}_1 = \bar{t} = \tilde{t}$ and $\hat{q}_1 = \bar{q} = \tilde{q}$ and the constraints become:

$$\begin{aligned} \tilde{t} - 2\bar{\theta}\tilde{q} &= 0 \\ \underline{t} - 2\underline{\theta}\underline{q} &= \hat{t}_2 - 2\hat{\theta}\hat{q}_2 \\ \hat{t}_2 - (\underline{\theta} + \bar{\theta})\hat{q}_2 &= \tilde{t} - (\underline{\theta} + \bar{\theta})\tilde{q} \end{aligned}$$

If we solve for the transfers, substitute in the objective function and then maximize with respect to \underline{q} , \hat{q}_2 and \tilde{q} we obtain the implicit definitions of the second part of the proposition. ■

Proof of Lemma 2. As in the case with no monitoring we want the grand contract to be delegation proof, i.e. $\Phi(\theta_1, \theta_2) = (\theta_1, \theta_2)$ but because of the monitoring the agent cannot misreport anymore the type of the second agent and the manipulation function boils down to a trivial version of the previous one $\Phi(\theta_1, \theta_2) = (\hat{\theta}_1, \theta_2)$. Given this and the fact that each agent can be only of two types, for each possible coalition, the coalition they could mimic, it is uniquely defined (for example $(\bar{\theta}, \underline{\theta})$ can pretend to be only $(\underline{\theta}, \underline{\theta})$). Therefore applying the same methodology of the proof of Lemma 1, the “coalition” incentive constraints are:

$$\begin{aligned} \underline{t} - 2\underline{\theta}\underline{q} &\geq \hat{t}_2 - 2\hat{\theta}\hat{q}_2 \\ \hat{t}_1 - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\hat{q}_1 &\geq \bar{t} - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\bar{q} \\ \hat{t}_2 - (\underline{\theta} + \bar{\theta})\hat{q}_2 &\geq \underline{t} - (\underline{\theta} + \bar{\theta})\underline{q} \\ \bar{t} - \left(2\bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\bar{q} &\geq \hat{t}_1 - \left(2\bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\hat{q}_1 \end{aligned}$$

As it is standard the only relevant constraints are the downward ones (namely the first two) that are binding at the optimum, the other two will be satisfied if the optimal schedule of output is monotonic. ■

Proof of Proposition 2. Considering the binding constraints (9), (10), (11) and (12) allows us to determine the incentive compatible and individually rational transfers, namely:

$$\underline{t} = 2\underline{\theta}\underline{q} + \Delta\theta\hat{q}_2 + \Delta\theta\bar{q}$$

$$\begin{aligned}\widehat{t}_2 &= (\underline{\theta} + \bar{\theta}) \widehat{q}_2 + \Delta\theta\bar{q} \\ \widehat{t}_1 &= \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right) \widehat{q}_1 + \frac{1-2\nu}{1-\nu}\Delta\theta\bar{q} \\ \bar{t} &= 2\bar{\theta}\bar{q}\end{aligned}$$

We can plug them into the principal's objective function and then maximize with respect to \underline{q} , \widehat{q}_1 , \widehat{q}_2 and \bar{q} , we then obtain the decreasing schedule of output of Proposition 2. ■

Proof of Lemma 3. The participation constraint and the coalition incentive compatibility constraints are the same as in the case where the principal can force the contractor to offer a separating subcontract. In addition there is the moral hazard constraint that should induce screening:

$$\nu U_1(\underline{\theta}, \underline{\theta}) + (1-\nu) U_1(\underline{\theta}, \bar{\theta}) \geq U_P^*(\underline{\theta}_1)$$

which rewrites as:

$$\nu(\underline{t} - 2\underline{\theta}\underline{q} - \Delta\theta\widehat{q}_1) + (1-\nu)(\widehat{t}_1 - (\underline{\theta} + \bar{\theta})\widehat{q}_1) \geq \widehat{t}_1 - (\underline{\theta} + \bar{\theta})\widehat{q}_1.$$

■

Proof of Proposition 3. The optimal contract, $GC = \{\underline{t}, \underline{q}, \widehat{t}_1, \widehat{q}_1, \widehat{t}_2, \widehat{q}_2, \bar{t}, \bar{q}\}$, now it is a solution to a program that maximizes the principal expected utility subject to the following constraints:

$$\begin{aligned}\widehat{t}_2 - (\underline{\theta} + \bar{\theta})\widehat{q}_2 - \Delta\theta\bar{q} &\geq 0 \\ \bar{t} - 2\bar{\theta}\bar{q} &\geq 0 \\ \underline{t} - 2\underline{\theta}\underline{q} &\geq \widehat{t}_2 - 2\underline{\theta}\widehat{q}_2\end{aligned}\tag{14}$$

$$\begin{aligned}\widehat{t}_1 - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\widehat{q}_1 &\geq \bar{t} - \left(\underline{\theta} + \bar{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\bar{q} \\ \underline{t} - 2\underline{\theta}\underline{q} &\geq \widehat{t}_1 - 2\underline{\theta}\widehat{q}_1\end{aligned}\tag{15}$$

where the first two are individual rationality constraints, the second pair are incentive compatibility constraints and the last constraint is the moral-hazard constraint.

It is standard to set the first two individual rationality constraints binding, the second incentive compatibility constraint is binding as well. The problem is to understand which one between (14) and (15) is binding. If we consider the optimal contract without this last constraint (the contract described in Prop.2) then at equilibrium, (15), the moral hazard constraint is not satisfied.

If instead we solve for the optimal contract neglecting (14) but with a binding

moral hazard constraint then the coalition incentive constraint rewrites as:

$$\widehat{q}_1 \geq (1 - v)\widehat{q}_2 + v\bar{q}$$

which is not satisfied. For this reason we define $\widehat{q}_1 = (1 - v)\widehat{q}_2 + v\bar{q}$ and solve for the other quantities in the standard way. Note that our solution slightly underestimates the equilibrium quantities except for \underline{q} . In order to be able to compute a solution we considered $S(\widehat{q}_1) = (1 - v)S(\widehat{q}_2) + vS(\bar{q})$ which is true only if \widehat{q}_1 is a random quantity (taking both values with certain probability) and not if it is a deterministic linear combination (since $S(\cdot)$ is a concave function then ■