DEPARTMENT OF ECONOMICS

UNIVERSITY OF MILAN - BICOCCA

WORKING PAPER SERIES

Monitoring Subcontracting in a Suppliers' Hierarchy

Michela Cella

No. 172 – July 2009

Dipartimento di Economia Politica
Università degli Studi di Milano - Bicocca
http://dipeco.economia.unimib.it

Monitoring Subcontracting in a Suppliers' Hierarchy*

Michela Cella†
University of Milan-Bicocca

This Version: July 2009

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 then 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, Andrea Prat and an anonymous referee. 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@unimib.it, Department of Economics, University Milan-Bicocca, P.za Ateneo Nuovo 1, 20126 Milano, Italy.

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 in the process of 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 Principle and its variations is that *ceteris paribus* a centralized structure always weakly

dominates a decentralized one[1].

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 to become informed about it because what happens at the subcontracting stage is private information to the agents.

Monitoring the communication between the contractor and the subcontractor would then 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 response[2].

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 looses the ability to manipulate the information about contracting costs for which, in the standard set-up with no monitoring, he receives and 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 will ensure 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.

[1] See Mookherjee and Tsumagari [2004] for a detailed and exhaustive analysis of the comparison between centralization and delegation.

[2] We can also think of a public register where the terms of the subcontract have to be recorded.

In other words the mark-up on contracting costs is now substituted by the incentives to screen the subcontractor's type, for some parameters 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][3] 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 [2006].

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. They show that centralization is dominated when collusion is taken into account and contracts are required to be anonymous (and therefore incomplete).

An analysis very similar to ours is carried out in Baron and Besanko [1992], where in a regulatory framework, they compare different organizational structures. They also consider costless monitoring in hierarchies 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, by making it dependent 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). We instead consider hierarchies where there are two productive agents and monitoring is done by the principal.

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.

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

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 5 concludes. All proofs are in the Appendix.

2 The Model

The principal P wants to buy a quantity $q \in \mathbb{R}^+$ 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. These inputs are perfect complements so that $q = q_1 = q_2$[4].

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 $\Theta_i = \Theta = \left\{\underline{\theta}, \overline{\theta}\right\}$, and $\Delta\theta = \overline{\theta} - \underline{\theta} > 0$. With probability ν (resp. $(1 - \nu)$) the agent is efficient, i.e. $\theta_i = \underline{\theta}$ (resp. inefficient, i.e. $\theta_i = \overline{\theta}$.).

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, with respect to the quantity, her revenue minus the monetary transfer to the first agent:

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

where $S(\cdot)$ is an increasing, strictly concave and twice continuously differentiable function that satisfies Inada conditions.

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 gives to the second agent at the subcontracting stage.

The second agent's utility is given by:

$$U_2 = y - \theta_2 q.$$

[4] In 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.

Both agents have reservation utility equal to zero.

If we had a centralized structure (where the principal directly contracts with each agent) we would obtain the following second best[5] quantities and rents:

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

In a centralized organization agents are treated symmetrically by the principal and obtain a positive informational rent only when they are efficient. In this case the principal maximizes his expected utility subject to the usual incentive compatibility and individual rationality constraints. Because each agent knows his own type, but not the one of the other agent, every constraint is at the interim stage. Expectations are taken according to the probability distribution of the agent's type. When of the efficient type each agent obtains an informational rent to ensure truthtelling. This rent takes into account that at the acceptance and reporting stage each agent knows his own type only. Production is downward distorted in those states of the world where an inefficient agent is present.

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 $q\left(\cdot\right)$ is total output, $t\left(\cdot\right)$ is the transfer from P to A_1 and $\left(\widehat{\theta}_1,\widehat{\theta}_2\right)$ are the types reported 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 message to be delivered to the principal[6] 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 A_2's reported type. The subcontract thus allows the agents to coordinate the reports to P, reallocate payments and possibly production assignments between themselves.

[5] Laffont and Martimort [1997 and 1998] show that this outcome is also collusion proof.

[6] This is a function that to any true pair of types assigns a pair of messages for the principal $\Phi:\Theta^2\rightarrow M_1\times M_2$. Then because of the Revelation Principle the relevant range for $\Phi\left(\theta_1,\theta_2\right)$ will be Θ^2.

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 $t\left(\overline{\theta},\overline{\theta}\right) = \overline{t}$; $t\left(\underline{\theta},\overline{\theta}\right) = \widehat{t}_1$; $t\left(\overline{\theta},\underline{\theta}\right) = \widehat{t}_2$; $t\left(\underline{\theta},\underline{\theta}\right) = \underline{t}$ and use a similar notation for $q\left(\cdot\right)$.

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 message 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.

The play of the game is such that the first agent decides on participating in the relationship with the principal only after receiving the report from the second agent. In other words, he will know the exact state of the world (i.e. both types) 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 aversion[7]. 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 centralization[8].

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

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

[8] This 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.

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 section [9].

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 message function and a transfer to the other agent, given the Grand Contract.

The following definition of a revealing subcontract will be useful throughout the paper.

Definition 1 *A revealing subcontract is a contract between A_1 and A_2 that reveals to A_2 the true type of A_1 at the offer stage.*

Contracting takes place under asymmetric information, so participation and incentive compatibility constraints for A_2 have to be considered when solving the following problem, $SC(\theta_1)$:

$$SC(\theta_1) = \begin{cases} \max\limits_{\substack{\Phi(\theta_1,\theta_i) \\ y(\theta_1,\theta_i)}} E_{\theta_2}[U_1(\theta_1)] = \nu\left(t\left(\Phi(\theta_1,\underline{\theta})\right) - y(\theta_1,\underline{\theta}) - \theta_1 q\left(\Phi(\theta_1,\underline{\theta})\right)\right) \\ \qquad\qquad + (1-\nu)\left(t\left(\Phi\left(\theta_1,\overline{\theta}\right)\right) - y\left(\theta_1,\overline{\theta}\right) - \theta_1 q\left(\Phi\left(\theta_1,\overline{\theta}\right)\right)\right) \\ \text{s.t.} \\ y\left(\theta_1,\overline{\theta}\right) - \overline{\theta} q\left(\Phi\left(\theta_1,\overline{\theta}\right)\right) = 0 \\ y(\theta_1,\underline{\theta}) - \underline{\theta} q\left(\Phi(\theta_1,\underline{\theta})\right) = y\left(\theta_1,\overline{\theta}\right) - \underline{\theta} q\left(\Phi\left(\theta_1,\overline{\theta}\right)\right) \end{cases} \tag{1}$$

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

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 type[10]. Rearranging the two binding constraints we obtain the transfers to the bottom agent:

$$y\left(\theta_1, \overline{\theta}\right) = \overline{\theta} q\left(\Phi\left(\theta_1, \overline{\theta}\right)\right) \tag{2}$$

$$y\left(\theta_1, \underline{\theta}\right) = \underline{\theta} q\left(\Phi\left(\theta_1, \underline{\theta}\right)\right) + \Delta\theta q\left(\Phi\left(\theta_1, \overline{\theta}\right)\right) \tag{3}$$

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

This means that the virtual cost for a unit $q\left(\Phi\left(\theta_1, \overline{\theta}\right)\right)$ is different from the true cost because it is the sum of production cost $\left(\overline{\theta}\right)$ and the informational rent that accrues to A_2 when he is efficient[11]. If we call $h_2\left(\theta_i\right)$ the virtual cost of a unit of q when the subcontractor is of type θ_i then we have:

$$h_2\left(\underline{\theta}\right) = \underline{\theta}$$
$$h_2\left(\overline{\theta}\right) = \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta$$

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*[12], a variant of the Revelation Principle,

[10] Since 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 and types are independent the principal cannot gain from concealing her private information. Therefore A_1 does not loose from making a revealing offer, i.e. offer a sub-contract which is dependent on his type.

[11] As in Mookherjee and Tsumagari [2004] we employ the term virtual cost because asymmetric information rests upon the costs of production. They are nothing more than the standard virtual types that take into account the informational rent that needs to be paid to efficient types to ensure truthful revelation.

[12] As it 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 agents, this of course cannot happen in delegation models where transfers are legitimate. For definition and application of Delegation Proofness and it's link with Collusion Proofness see Laffont and Martimort [1998] and Faure-Grimaud, Laffont and Martimort [2003].

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 message function is equal to the identity function ($\Phi\left(\theta_1, \widetilde{\theta}_2\right) = \left(\theta_1, \widetilde{\theta}_2\right)$) because truthtelling is the optimal strategy.

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* $\underline{q} \geq \widehat{q}_2 \geq \widehat{q}_1 \geq \overline{q}$ *and the following incentive compatibility constraints are satisfied:*

$$\underline{t} - 2\underline{\theta}\underline{q} \geq \widehat{t}_2 - 2\underline{\theta}\widehat{q}_2 \tag{4}$$

$$\widehat{t}_2 - \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_2 \geq \widehat{t}_1 - \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_1 \tag{5}$$

$$\widehat{t}_1 - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\widehat{q}_1 \geq \overline{t} - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\overline{q} \tag{6}$$

The above constraints give the conditions that the transfers to the contractor have to satisfy to obtain a truthful report in the Grand Contract in each state of the world. We know from standard results in mechanism design that if the schedule of output is monotone we have to take care only of adjacent upward constraints.

In our setting the cost of a particular pair of productive agents is the sum of two elements: the cost of the contractor and the virtual cost of the subcontractor. The virtual cost of an efficient subcontractor coincide with his cost, while in the case of an inefficient on the virtual cost is higher because of the informational rent paid at the subcontracting stage.

As a consequence the total cost of mixed pairs are not equal, the cost of a pair that includes an inefficient second agent is higher. If we impose monotonicity on the schedule of output we ensure that the mixed coalition with higher total costs will not want to mimic the other.

In synthesis, in our framework we have four levels of total cost of production and the contractor, A_1, has to be given incentives to truthfully report them. He therefore obtains an informational rent that constitutes an additional mark-up on the subcontracting costs and that is the source of the higher costs of delegation.

The principal then optimally trades off rents and efficiency offering a schedule of output that is more downward distorted than the second best one, the one that is optimal under centralized contracting.

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

$$\begin{aligned}\max E_{\theta_1,\theta_2}[W] &= \nu^2\left(S\left(\underline{q}\right)-\underline{t}\right)+\nu\left(1-\nu\right)\left(S\left(\widehat{q}_1\right)-\widehat{t}_1\right)+ \\ &\quad +\nu\left(1-\nu\right)\left(S\left(\widehat{q}_2\right)-\widehat{t}_2\right)+\left(1-\nu\right)^2\left(S\left(\overline{q}\right)-\overline{t}\right)\end{aligned} \tag{7}$$

Subject to incentive compatibility constraints (4-6) and the following set of ex-post individual rationality constraints:

$$t_{ij}-\theta_i q_{ij}-y_{ij}\geq 0 \quad \forall i,j=1,2. \tag{8}$$

Conditional on the optimal schedule of output being monotone and the constraints (4)-(6) being satisfied we can restrict attention to the participation constraints of a pair of two inefficient agents:

$$\overline{t}-2\overline{\theta}\overline{q}=0 \tag{9}$$

These considerations simplify the optimization problem, from the constraints we obtain the incentive feasible transfers which allow to solve for the optimal contract, described 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* $\underline{q}>\widehat{q}_2>\widehat{q}_1>\overline{q}$ *where the prescribed quantities are implicitly defined by:*
    - $S'\left(\underline{q}\right)=2\underline{\theta}$
    - $S'\left(\widehat{q}_2\right)=\overline{\theta}+\underline{\theta}+\frac{\nu}{1-\nu}\Delta\theta$
    - $S'\left(\widehat{q}_1\right)=\underline{\theta}+\overline{\theta}+\frac{\nu(2-\nu)}{(1-\nu)^2}\Delta\theta$
    - $S'\left(\overline{q}\right)=2\overline{\theta}+\frac{\nu(2-\nu)(1-2\nu)}{(1-\nu)^3}\Delta\theta$
  - *The ex-post agents' payoffs are the following:*
    - $U_1\left(\underline{\theta},\underline{\theta}\right)=\Delta\theta\left(\widehat{q}_2-\widehat{q}_1\right)+\frac{\nu}{1-\nu}\Delta\theta\widehat{q}_1+\frac{1-2\nu}{(1-\nu)}\Delta\theta\overline{q}$
    - $U_1\left(\underline{\theta},\overline{\theta}\right)=\Delta\theta\overline{q}+\frac{\nu}{1-\nu}\Delta\theta\left(\widehat{q}_1-\overline{q}\right)$
    - $U_1\left(\overline{\theta},\underline{\theta}\right)=\frac{\nu}{1-\nu}\Delta\theta\left(\widehat{q}_1-\overline{q}\right)$
    - $U_1\left(\overline{\theta},\overline{\theta}\right)=0$
    - $U_2\left(\underline{\theta},\underline{\theta}\right)=\Delta\theta\widehat{q}_1$
    - $U_2\left(\overline{\theta},\underline{\theta}\right)=\Delta\theta\overline{q}$
    - $U_2\left(\theta_i,\overline{\theta}\right)=0$
- *for* $\nu\geq\nu^*$

- *It implements a decreasing schedule of outputs with some bunching* $\underline{q} > \widehat{q}_2 > \widetilde{q} = \widehat{q}_1 = \overline{q}$ *where the prescribed quantities are implicitly defined by:*
  - $S'\left(\underline{q}\right) = 2\underline{\theta}$
  - $S'\left(\widehat{q}_2\right) = \overline{\theta} + \underline{\theta} + \frac{\nu}{1-\nu}\Delta\theta$
  - $S'\left(\widetilde{q}\right) = 2\overline{\theta} + \frac{\nu}{(1-\nu)}\Delta\theta$
- *The ex-post agents' payoffs are the following:*
  - $U_1\left(\underline{\theta}, \underline{\theta}\right) = \Delta\theta\widehat{q}_2$
  - $U_1\left(\underline{\theta}, \overline{\theta}\right) = \Delta\theta\widetilde{q}$
  - $U_1\left(\overline{\theta}, \underline{\theta}\right) = U_1\left(\overline{\theta}, \overline{\theta}\right) = 0$
  - $U_2\left(\underline{\theta}, \underline{\theta}\right) = U_2\left(\overline{\theta}, \underline{\theta}\right) = \Delta\theta\widetilde{q}$
  - $U_2\left(\theta_i, \overline{\theta}\right) = 0$

The above contract requires quantities that are more downward distorted than those of the second best one. The amount of informational rent paid to the agents is higher than that under centralization 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 with an inefficient second agent, 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 a transfer which includes a reimbursement of the virtual cost plus an informational rent to reveal it. This rent is obviously higher than the one the first agent would get if he just had to report his own cost.

If we look at equilibrium payoffs, we can see that the bottom agent is treated as in the second best contract: he receives a positive rent only when he is efficient, and at the interim stage (that is before knowing the type of the first agent) they are equal. It is instead different what happens to the informational rent of the first agent. When $\nu < \nu^*$, he obtains a positive rent also when he is inefficient and paired with an efficient second agent, this is due to the double informative advantage at the acceptance and reporting stage. When $\nu \geq \nu^*$ the probability of facing an efficient agent increases and the screening of a coalition of the $\left(\underline{\theta}, \overline{\theta}\right)$ type proves so costly that the principal gains by bunching the quantities which involve an inefficient second agent. In this case an inefficient first agent will get no rent irrespectively of the type of the other agent, exactly like in the second best. When he is efficient the first agent obtains a rent which is higher than the second best at the interim stage.

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.

We can imagine the principal having access to some public register where the agreed subcontract must be recorded or as if the principal sent a person of trust to be present but silent at the subcontract negotiation 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 before it will be reported by the middle agent.

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 his own type. This will leave an efficient first agent with a payoff that does not exceed his reservation utility, clearly worse off than without the monitoring.

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 A_1 for the revelation of A_2's information, therefore saving on what we called the "true" cost of delegation.

It should be clear 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. The monitoring takes place when he accepts the subcontract and reveals his information that is after the incentives for the truthful revelation of his type have been designed. It is the contractor who is damaged the most by the monitoring, he could be left with no rent at all in any state of the world.

Our aim is to study the reaction of the contractor when he moves and offers the subcontract; he might change his offer in an attempt to conceal some information from the principal and get some informational rent back.

The first and most obvious reaction would be to conceal his own type when offering the subcontract, as we will see 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.

To condition a contract on the type of the offering party without revealing it the agent has to offer a menu of quantities and transfers that includes the optimal ones for each of the possible states of the world. In our specific case this means offering four pairs of quantities and transfers, each one designed for one of the possible pairs of producing agents, even though at the time the offer is made the contractor knows

that two of them will never be implemented[13]. This is just an application of a result by Maskin and Tirole [1990] that show that, in a world of private values[14] and with quasilinear utility functions, the agent receiving the offer is indifferent between being informed about the principal's type or not. In our case there is no advantage in doing so vis-a-vis the subcontractor, the gain, in fact, 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[15], the ones that satisfy ex-post constraints. In other words the transfers included in a not-revealing subcontract offer will be:

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

$$SC\left(\theta_1, \overline{\theta}\right) = \left\{y\left(\theta_1, \overline{\theta}\right), \Phi\left(\theta_1, \overline{\theta}\right); \quad \theta_1 \in \Theta\right\}$$

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 and therefore which transfer he will receive.

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.

At this stage the principal/monitor no longer learns the type of the contractor but can nonetheless still observe the report done by the subcontractor about his type. This means that at the subsequent 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, where each pair is deigned for a particular type of A_2.

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*

[13] In the real world one can imagine an overly complicated contract being offered, useless conditions that the offering party knows will never be applied.

[14] We are in a private values framework when the type of the principal (the offering party) is not an argument of the agent's utility function.

[15] The degress of freedom when solving the maximization problem allow us to choose the same transfers of the benchmark case, thus ensuring the subcontractor indifference.

($\underline{q} \geq \widehat{q}_2$ and $\widehat{q}_1 \geq \overline{q}$)and the following constraints are satisfied:

$$\underline{t} - 2\underline{\theta}\underline{q} \geq \widehat{t}_2 - 2\underline{\theta}\widehat{q}_2 \tag{10}$$

$$\widehat{t}_1 - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\widehat{q}_1 \geq \overline{t} - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\overline{q} \tag{11}$$

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 is informed about 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. This implies that A_1 must be given incentive to report truthfully just his own type.

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 choice of the relevant 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:

$$\widehat{t}_2 - \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_2 - \Delta\theta\overline{q} = 0 \tag{12}$$

$$\overline{t} - 2\overline{\theta}\overline{q} = 0 \tag{13}$$

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 should now be 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 $\underline{q} > \widehat{q} > \overline{q}$ (where $\widehat{q} = \widehat{q}_1 = \widehat{q}_2$) implicitly defined by:*
  - $S'\left(\underline{q}\right) = 2\underline{\theta}$

- $S'(\widehat{q}) = \left(\underline{\theta} + \overline{\theta}\right) + \frac{\nu}{1-\nu}\Delta\theta$
- $S'(\overline{q}) = 2\overline{\theta} + \frac{2\nu}{1-\nu}\Delta\theta$

- *The ex-post agents' payoffs are the following:*

  - $U_1(\underline{\theta}, \underline{\theta}) = \Delta\theta\overline{q}$
  - $U_1\left(\underline{\theta}, \overline{\theta}\right) = \frac{\nu}{1-\nu}\Delta\theta\widehat{q} + .\frac{1-2\nu}{(1-\nu)}\Delta\theta\overline{q}$
  - $U_1\left(\overline{\theta}, \theta_i\right) = 0$
  - $U_2(\underline{\theta}, \underline{\theta}) = \Delta\theta\widehat{q}$
  - $U_2\left(\overline{\theta}, \underline{\theta}\right) = \Delta\theta\overline{q}$
  - $U_2\left(\theta_i, \overline{\theta}\right) = 0$

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.[16] 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.

If we look more carefully 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 across states of the world. These rents are lower than the ones the agents earn in our benchmark model of delegation with no monitoring.

Looking at the ex-post payoffs of the agents we can see why the principal benefits from monitoring. The gain comes from a reduction of the payments made to the prime contractor who earns lower rents than when monitoring was not possible. The subcontractor at the bottom of the hierarchy is instead unaffected by the monitoring because he receives the same incentives for a truthful report to the contractor whether monitoring happens or not.

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

It is, in fact, only after the subcontract has been offered and accepted that the principal gets to know A_2's type. The gains from monitoring accrue to the principal because he will not pay the (above mentioned) markup on subcontracting costs to A_1. She knows the type of A_2 and is not giving A_2 any incentive for a truthful report. The contractor obtains an informational rent only to reveal his own type and will be only reimbursed the subcontracting costs.

The obvious problem with the optimal contract of Proposition 2 is that knowing that he cannot extract rent from the revelation of A_2's type, A_1 may prefer to ignore A_2's type. To do that, he may offer a subcontract that does not screen A_2's type. If the principal's offer cannot be contingent on the type of subcontract, this deviation is feasible and profitable.

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-contract. This can be achieved through a pooling subcontract , defined below, that does not require a report and that is independent from A_2's type.

Definition 2 *A pooling subcontract is a contract between A_1 and A_2 that does not separate the types of A_2 and hence does not require any report from the second agent. More precisely it is composed by a message function $\Phi(\theta_1, \theta_2)$ and the following set of transfers:*

$$y(\theta_1, \theta_2) = \overline{\theta} q(\Phi(\theta_1, \theta_2)) \qquad \forall \theta_1, \theta_2.$$

In a pooling subcontract the first agent will offer a set of transfers to the second one as if the latter was always inefficient, 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:

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

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

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 the were dependent on both types. Actually in this case, the message space for the first agent when reporting to the principal is as large as in the benchmark case with no monitoring, earlier in this section, because of the monitoring, A_1 was restricted to the message space $\left\{\widehat{\theta}_i, \theta_2\right\}$ (he had to report the true θ_2). Now the message space is in fact $\left\{\widehat{\theta}_i, \widehat{\theta}_j\right\}$ (with $i, j = 1, 2$) but the pooling subcontract implies

that $\widehat{\theta}_j = \overline{\theta}$ always. As a consequence of this pooling contract the message function is reduced once again to a function of one variable: $\Phi(\theta_1, \theta_2) = \left(\widehat{\theta}_1, \overline{\theta}\right)$.

Since the agent might in fact prefer to offer a pooling subcontract, he needs to be given incentive to pick a screening subcontract, given our assumption that the subcontract cannot be contracted upon. This is in line with our benchmark model where in fact the Grand Contract does not require a screening subcontract between the agents. The contractor screens because it's in his interest to do so.

Our goal is to study strategic interactions among members of a contracting hierarchy so allowing the principal to just impose a particular form of subcontract upon the contractor and greatly reduce his autonomy would not seem consistent with our scope. We have deliberately chosen a framework that gives the most freedom to the first agent, our choice of timing (with acceptance of the Grand Contract after receiving the second agent's report) is another step in this direction.

We need to study which subcontract offer is optimal for A_1. We require subgame perfection and 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 $GC = \left\{\underline{t}, \underline{q}, \widehat{t}_1, \widehat{q}_1, \widehat{t}_2, \widehat{q}_2, \overline{t}, \overline{q}\right\}$ the contractor will choose the type of subcontract that maximizes his expected utility.

The expected utility for a contractor of type θ_1 of offering a separating subcontract is:

$$\begin{aligned} E_{\theta_2}\left[U_1(\theta_1)\right] &= \nu\left(t\left(\Phi(\theta_1, \underline{\theta})\right) - \underline{\theta} q\left(\Phi(\theta_1, \underline{\theta})\right) - \Delta\theta q\left(\Phi\left(\theta_1, \overline{\theta}\right)\right) - \theta_1 q\left(\Phi(\theta_1, \underline{\theta})\right)\right) \\ &\quad (1-\nu)\left(t\left(\Phi\left(\theta_1, \overline{\theta}\right)\right) - \overline{\theta} q\left(\Phi\left(\theta_1, \overline{\theta}\right)\right) - \theta_1 q\left(\Phi\left(\theta_1, \overline{\theta}\right)\right)\right) \end{aligned} \tag{14}$$

while the expected utility of offering a pooling subcontract is:

$$U_P(\theta_1) = t\left(\Phi\left(\theta_1, \overline{\theta}\right)\right) - \left(\overline{\theta} + \theta_1\right) q\left(\Phi\left(\theta_1, \overline{\theta}\right)\right) \ . \tag{15}$$

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:

$$\nu U_1(\underline{\theta}, \underline{\theta}) + (1-\nu) U_1\left(\underline{\theta}, \overline{\theta}\right) \geq U_P^*(\underline{\theta}_1) \tag{16}$$

where $U_1(\underline{\theta}, \underline{\theta})$ and $U_1\left(\underline{\theta}, \overline{\theta}\right)$ are the rents earned by an efficient first agent, who

is paired with an efficient and an inefficient second agent respectively, when he offers a separating subcontract and truthfully reports to the principal. 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:

$$U_P^*(\underline{\theta}_1) = \max_{\Phi} t\left(\Phi\left(\underline{\theta}_1, \overline{\theta}\right)\right) - \left(\overline{\theta} + \underline{\theta}_1\right) q\left(\Phi\left(\underline{\theta}_1, \overline{\theta}\right)\right)$$

We 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.

Constraint (16) is in fact a moral hazard constraint. At the contract design stage the principal has to give incentives to the first agent to do her preferred action which, in this case, is offering a screening contract. When 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 so 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 Lemma is of some help in this direction.

Lemma 3 *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:*

$$U_P^*(\underline{\theta}_1) = \widehat{t}_1 - \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_1.$$

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 sub-contract offer.

Lemma 4 *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 subcontract if the following constraints are satisfied:*

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

and the output schedule is such that $\underline{q} \geq \widehat{q}_2$ and $\widehat{q}_1 \geq \overline{q}$.

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 level[17]. 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 characterize 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 cannot force A_1 to offer a screening subcontract the optimal grand contract has the following characteristics:*

- *It implements a decreasing schedule of output $\underline{q} > \widehat{q}_2 > \widehat{q}_1 > \overline{q}$ (where $\widehat{q}_1 = (1-\nu)\,\widehat{q}_2 + \upsilon\overline{q}$) with the following properties:*
  - $S'\left(\underline{q}\right) = 2\underline{\theta}$
  - $\widehat{q}_2 > \widehat{q}_2^{\circ}$ *where* $S'\left(\widehat{q}_2^{\circ}\right) = \left(\underline{\theta} + \overline{\theta}\right) + \frac{\nu}{1-\nu}\Delta\theta$
  - $\overline{q} < \overline{q}^{\circ}$ *where* $S'\left(\overline{q}^{\circ}\right) = 2\overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta + \frac{\upsilon(1-\nu)}{1-\nu+\upsilon^2}\Delta\theta$
- *The ex-post agents' payoffs are the following:*
  - $U_1\left(\underline{\theta}, \theta_i\right) = \Delta\theta\left[\nu\widehat{q}_2 + (1-\nu)\,\overline{q}\right]$
  - $U_1\left(\overline{\theta}, \theta_i\right) = 0$
  - $U_2\left(\underline{\theta}, \underline{\theta}\right) = \Delta\theta\widehat{q}_1 = \Delta\theta\left[(1-\nu)\,\widehat{q}_2 + \upsilon\overline{q}\right]$
  - $U_2\left(\overline{\theta}, \underline{\theta}\right) = \Delta\theta\overline{q}$
  - $U_2\left(\theta_i, \overline{\theta}\right) = 0$

The optimal contract when the agent is free to choose the type of subcontract and the principal is monitoring generates an equilibrium which, from the point of view of the principal, is worse than the second best outcome. 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 before, in the benchmark case without monitoring, the contractor had full flexibility on what to tell the principal about the report from the subcontractor, now he has, at least, the freedom to ask for a report.

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

The quantities are slightly more distorted than the second best ones and because of the newly introduced moral hazard constraint there is some form of bunching, $\widehat{q}_1$ is not optimally determined but it is an average of $\widehat{q}_2$ and $\overline{q}$. The first agent's informational rent are higher than in the second best.

Choosing a specific functional form allows us to find exact values for the optimal quantities.

Corollary 1 *If the principal gross surplus function is* $S(q) = Kq - \frac{q^2}{2}$ *(with* K *large enough):*

- *the optimal quantities are:*
  - $\underline{q} = K - 2\underline{\theta}$
  - $\widehat{q}_2 = K - \left(\underline{\theta} + \overline{\theta}\right) - \frac{\nu}{1-\nu}\Delta\theta$
  - $\widehat{q}_1 = K - \left(\underline{\theta} + \overline{\theta}\right) - \frac{\upsilon\left(2-2\nu+\nu^2\right)}{(1-\nu)(1-\nu+\upsilon^2)}$
  - - $\overline{q} = K - 2\overline{\theta} - \frac{\nu}{1-\nu}\Delta\theta - \frac{\upsilon(1-\nu)}{1-\nu+\upsilon^2}\Delta\theta$
- *with* $\underline{q} > \widehat{q}_2 > \widehat{q}_1 > \overline{q}$,
- A_1*'s payoffs are the same of Proposition 3.*

We are now in the condition to study how the contractor's reaction to the monitoring can influence the decision of the principal to monitor at all.

Proposition 4 *For* $\nu < \nu^*$ *the principal prefers to monitor and give incentives to the contractor for a screening subcontract. For* $\nu \geq \nu^*$ *the principal prefers not to monitor.*

For a low ν the contract of Proposition 3 earns a higher net surplus to the principal, rents for the agents are lower and quantities, as a consequence, less distorted away from the first best. Avoiding the double marginalization of rents through monitoring is the key driving force. While monotonicity is ensured by construction ($\widehat{q}_1 = (1-\nu)\,\widehat{q}_2 + \upsilon\overline{q}$), when ν, the probability of facing an efficient agent, is higher expected rents grow substantially. As a consequence the principal prefers to avoid monitoring and bunch some types to save on rents.

It is now evident that the reaction by the contractor to the monitoring by the principal causes extra-costs. As long as the probability of dealing with an efficient agent is low enough the monitoring and incentives to screen are profitable. When this probability increases, screening all the possible agents' pairs becomes too costly, whether it is achieved in the "standard" way (no monitoring) or via monitoring coupled with the incentives for a screening subcontract.

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, to the effect that in some cases the principal prefers to avoid monitoring, although costless.

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. With two types a subcontract will either be fully screening or fully pooling, with more types and form of semi-separating is possible. We conjecture, but have not proven, that the qualitative results would not change if the type space was richer. When the number of types is greater the costs of screening all the possible pairs would increase steeply and so optimal contracts would quite surely involve some form of bunching. At the same time the costs of not having a screening sub-contract, and consider all agents of the lower type, are likely to increase.

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 given the independence of the contractor and the timing of the game.

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. (2006) "Delegation and Contracting Hierarchies: An Overview", *Journal of Economic Literature*, 44:367-390

[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. The private information in the hands of A_1 when he deals with the principal is given by his cost plus the virtual cost of A_2. From the point of view of the principal the total costs of a pair of agents are given by the sum of the first agent's cost plus the virtual costs of the second one.

Total costs, for each pair, are given by:
$(\underline{\theta},\underline{\theta})$: $\underline{\theta}+\underline{\theta}$
$(\overline{\theta},\underline{\theta})$: $\overline{\theta}+\underline{\theta}$
$(\underline{\theta},\overline{\theta})$: $\underline{\theta}+\overline{\theta}+\frac{\nu}{1-v}\Delta\theta$
$(\overline{\theta},\overline{\theta})$: $\overline{\theta}+\overline{\theta}+\frac{\nu}{1-v}\Delta\theta$.

Incentive constraints for a pair of agents of type (θ_i,θ_j) then take the form:

$$t_{ij}-(\theta_i+h(\theta_j))\,q_{ij}\geq t_{kl}-(\theta_i+h(\theta_j))\,q_{kl}$$

$\forall i,j,k,l=1,2$ and $i\neq k$, $j\neq l$ and where $h(\theta_j)$ is the virtual cost. Adding monotonicity with respect to virtual costs allows us to restrict attention to local upward incentive constraints. ■

Proof of Proposition 1. Given the binding constraints we can manipulate them and obtain the incentive compatible and individually rational transfers:
$\underline{t}=2\underline{\theta}\underline{q}+\Delta\theta\widehat{q}_2+\frac{\nu}{1-\nu}\Delta\theta\widehat{q}_1+\frac{1-2\nu}{1-\nu}\Delta\theta\overline{q}$
$\widehat{t}_2=\left(\underline{\theta}+\overline{\theta}\right)\widehat{q}_2+\frac{\nu}{1-\nu}\Delta\theta\widehat{q}_1+\frac{1-2\nu}{1-\nu}\Delta\theta\overline{q}$
$\widehat{t}_1=\left(\underline{\theta}+\overline{\theta}+\frac{\nu}{1-\nu}\Delta\theta\right)\widehat{q}_1+\frac{1-2\nu}{1-\nu}\Delta\theta\overline{q}$
$\overline{t}=2\overline{\theta}\overline{q}$

Substitute them in the principal's objective function and then maximize with respect to $\underline{q}$, $\widehat{q}_1$, $\widehat{q}_2$ and $\overline{q}$, we then obtain the decreasing schedule of output in the first part of Proposition 1.

We need to check that monotonicity with respect to virtual costs is satisfied, we find that $\widehat{q}_1>\overline{q}$ is true only when:

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

The above holds 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 $\widehat{t}_1=\overline{t}=\widetilde{t}$ and $\widehat{q}_1=\overline{q}=\widetilde{q}$ and the

constraints become:

$$\begin{aligned} \widetilde{t} - 2\overline{\theta}\widetilde{q} &= 0 \\ \underline{t} - 2\underline{\theta}\underline{q} &= \widehat{t}_2 - 2\underline{\theta}\widehat{q}_2 \\ \widehat{t}_2 - \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_2 &= \widetilde{t} - \left(\underline{\theta} + \overline{\theta}\right)\widetilde{q} \end{aligned}$$

If we solve for the transfers, substitute in the objective function and then maximize with respect to $\underline{q}$, , $\widehat{q}_2$ and $\widetilde{q}$ 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 message function boils down to a trivial version of the previous one $\Phi(\theta_1, \theta_2) = \left(\widehat{\theta}_1, \theta_2\right)$, where only θ_1 is truly reported. Given this and the fact that each agent can be only of two types, for each possible pair, the pair they could mimic is uniquely defined (as an example a pair $(\underline{\theta}, \underline{\theta})$ can only pretend to be $(\overline{\theta}, \underline{\theta})$). In other words, if the quantities are monotonic with respect to virtual cost, constraints need to take care only of the incentive of an efficient A_1 to upward distort his report. All these considerations allow us to restrict attention to the following constraints:

$\underline{t} - 2\underline{\theta}\underline{q} \geq \widehat{t}_2 - 2\underline{\theta}\widehat{q}_2$

$\widehat{t}_1 - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\widehat{q}_1 \geq \overline{t} - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\overline{q}$ ■

Proof of Proposition 2. Considering the binding constraints (10), (11), (12) and (13) allows us to determine the incentive compatible and individually rational transfers, namely:

$\underline{t} = 2\underline{\theta}\underline{q} + \Delta\theta\widehat{q}_2 + \Delta\theta\overline{q}$

$\widehat{t}_2 = \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_2 + \Delta\theta\overline{q}$

$\widehat{t}_1 = \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\widehat{q}_1 + \frac{1-2\nu}{1-\nu}\Delta\theta\overline{q}$

$\overline{t} = 2\overline{\theta}\overline{q}$

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

Proof of Lemma 3. When offering a pooling subcontract A_1 will truthfully report his type if the following constraint is satisfied:

$$\widehat{t}_1 - \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_1 \geq \overline{t} - \left(\underline{\theta} + \overline{\theta}\right)\overline{q}$$

Instead, if the subcontract is separating, the Grand Contract is delegation proof if

the following two constraints are satisfied:

$$\underline{t} - 2\underline{\theta}\underline{q} \geq \widehat{t}_2 - 2\underline{\theta}\widehat{q}_2$$

$$\widehat{t}_1 - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\widehat{q}_1 \geq \overline{t} - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\overline{q}$$

If we subtract $\frac{\nu}{1-\nu}\Delta\theta\widehat{q}_1$ from both sides of the inequality in the first constraint we obtain the latter. It is then evident that the first one is satisfied whenever the last two are. ■

Proof of Lemma 4. 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, due to the monitoring in fact the first agent cannot misreport the second agent's type. In addition there is the moral hazard constraint that should induce screening:

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

that, after one substitutes for the transfers, rewrites as:

$$\nu\left(\underline{t} - 2\underline{\theta}\underline{q} - \Delta\theta\widehat{q}_1\right) + (1-\nu)\left(\widehat{t}_1 - \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_1\right) \geq \widehat{t}_1 - \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_1$$

that simplify to:

$$\underline{t} - 2\underline{\theta}\underline{q} \geq \widehat{t}_1 - 2\underline{\theta}\widehat{q}_1$$

■

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

$$\begin{aligned}
\widehat{t}_2 - \left(\underline{\theta} + \overline{\theta}\right)\widehat{q}_2 - \Delta\theta\overline{q} &\geq 0 \\
\overline{t} - 2\overline{\theta}\overline{q} &\geq 0 \\
\underline{t} - 2\underline{\theta}\underline{q} &\geq \widehat{t}_2 - 2\underline{\theta}\widehat{q}_2 \qquad (17) \\
\widehat{t}_1 - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\widehat{q}_1 &\geq \overline{t} - \left(\underline{\theta} + \overline{\theta} + \frac{\nu}{1-\nu}\Delta\theta\right)\overline{q} \\
\underline{t} - 2\underline{\theta}\underline{q} &\geq \widehat{t}_1 - 2\underline{\theta}\widehat{q}_1 \qquad (18) \\
\underline{q} &\geq \widehat{q}_2 \text{ and } \widehat{q}_1 \geq \overline{q} \qquad (19)
\end{aligned}$$

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

straint.

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 (17) and (18) is binding. If we consider the optimal contract without this last constraint (the contract described in Prop.2) then at equilibrium (18), the moral hazard constraint, is not satisfied.

If instead we solve for the optimal contract neglecting (17) but with a binding moral hazard constraint then the coalition incentive constraint rewrites as:

$$\widehat{q}_1 \geq (1-\upsilon)\,\widehat{q}_2 + \upsilon\overline{q}$$

which is not satisfied. For this reason we set it binding and define $\widehat{q}_1 = (1-\upsilon)\,\widehat{q}_2 + \upsilon\overline{q}$, knowing that $\widehat{q}_1$ will not be optimally determined. Since $S(\cdot)$ is a concave function we cannot solve explicitly without choosing a particular (and very simple) functional form. We then solve for the quantities that would be optimal if $S(\cdot)$ was linear, that is when:

$$S'(\widehat{q}_1) = (1-\upsilon)\,S'(\widehat{q}_2) + \upsilon S'(\overline{q})$$

This allows us to find exact solutions for $\widehat{q}_2^{\circ}$ and $\overline{q}^{\circ}$ in the standard way, then knowing that:

$$S'((1-\upsilon)\,\widehat{q}_2 + \upsilon\overline{q}) > S'(\widehat{q}_2)$$

and

$$S'((1-\upsilon)\,\widehat{q}_2 + \upsilon\overline{q}) < S'(\overline{q})$$

we can say that the optimal quantities satisfy $\widehat{q}_2 > \widehat{q}_2^{\circ}$ and $\overline{q} < \overline{q}^{\circ}$. This guarantees that the two monotonicity constraints are satisfied. ■

Proof of Proposition 4. When $\nu < \nu^*$the optimal quantities with monitoring and the moral hazard constraints are higher for any type of agent's pairs. This guarantees a higher net surplus to the principal.

When $\nu \geq \nu^*$we observe that $\overline{q} < \widetilde{q} < \widehat{q}_1$. Expected rents are though higher in the monitoring case therefore the expected net surplus is higher in the benchmark case of no monitoring. ■