Optimal assignment of principalship in teams

Shuhe Li\textsuperscript{a,\#}, Weiying Zhang\textsuperscript{b}

\textsuperscript{a} Department of Economics and Finance, Faculty of Business, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
\textsuperscript{b} Guanghua School of Management, Peking University, Beijing 100871, China

Received 20 August 1998; received in revised form 23 March 1999; accepted 31 March 1999

Abstract

This paper investigates the optimal assignment of principalship in teams. We formalize a theory of the firm pioneered by Alchian and Demsetz [Am. Econ. Rev. 62 (1972) 777] using an approach in the spirit of Grossman and Hart [J. Pol. Econ. 94 (1986) 691]. Our theory makes a distinction between production effort and monitoring effort. When it is costly to sign contracts on efforts, it may be optimal to let one party purchase the rights to monitor and to direct, and claim full residual. Principalship is the purchase of these rights. These rights are limited residual rights of control over actions. We have confirmed Alchian and Demsetz’s celebrated conjecture that residual claims should match residual control. We have also explored how the optimal assignment of principalship and partnership depends on the interaction between each member’s importance in production, the effectiveness of monitoring and the degree of teamwork. © 2001 Elsevier Science B.V. All rights reserved.

JEL classification: D2; D8; L2

Keywords: Principal agent; Partnership; Monitoring; Team

1. Introduction

1.1. General introduction

The standard agency theory of hidden actions focuses on the optimal design of an incentive scheme by a party whose principalship is exogenously given and who does not contribute production effort directly, i.e. he is an ‘outsider’. In this paper, we consider a...
team production in which all members have production inputs and thus all are ‘insiders’. We address the questions of when principalship Pareto dominates partnership and who will emerge as the principal. In our proposed set-up, we can deal with the problems of designing an incentive scheme and assigning principalship simultaneously. We want to emphasize that in certain circumstances, optimal assignment of principalship may have primary order effects on efficiency, whereas the design of an incentive scheme may have only secondary order effects on efficiency. For instance, consider a team of two members: one works in the dark and another works in the light. The worker-in-the-dark can easily monitor the worker-in-the-light, and the worker-in-the-light can hardly monitor the worker-in-the-dark, no matter what incentive scheme is adopted. It is optimal to assign principal/monitorship to the worker-in-the-dark, and the design of an incentive scheme may be less important.¹

A theory of endogenous principalship shall identify the determinants of optimal assignment of principalship. Possible determinants include the distributions of physical capital, human capital, information and risk attitudes of all parties involved as well as the nature of production. See Zhang (1994) for more discussions along this line. As a first step, in this paper we investigate how the distributions of advantages in activities, hidden actions and the nature of teamwork affect the optimal assignment of principalship in teams.

We shall formalize a theory of the firm pioneered by Alchian and Demsetz (1972) using an approach in the spirit of Grossman and Hart (1986). Our theory of principalship in teams makes a distinction between two types of efforts: production effort and monitoring effort. When it is costly to design and sign contracts on efforts, it may be optimal to let one party purchase the rights to monitor and to direct others, and claim full residual. Principalship is the purchase of these rights. These rights are limited residual rights of control over actions. They are limited by the right to withdraw and the hidden action of the agent. We find that principalship is the primary incentive mechanism in teams. The optimal assignment of principalship depends on interactions between each member’s relative importance in production, the effectiveness of monitoring and the degree of teamwork. As expected, a team member is more likely to hold principalship when he has advantages in both production and monitoring. A less intuitive finding is that a higher degree of teamwork tends to call for the member who has comparative advantage in monitoring to hold principalship. Partnership is desirable if all the members are ‘somewhat’ important in production and ineffective in monitoring.

As Alchian and Demsetz (1972) pointed out, the essence of the classical firm is team production of several input owners. Nonseparability of the products of these input owners raises the cost of measuring the individual contributions of those resources. To reduce shirking, it is necessary for a team member to specialize in monitoring others. To discipline team members, the monitor must have the authority to design job assignments and rewards. But the monitor’s authority is subject to continuous renegotiation, with each member, of the terms that must be acceptable to both the monitor and the member. To induce self-monitoring, the monitor holds the right to claim full residual, i.e. residual claims should match residual control rights.

¹ A related real world example is a garment firm with a fashion designer and a sewer. Since it is much easier to direct and to measure the marginal productivity of the sewer (worker-in-the-light) than that of the fashion designer (worker-in-the-dark), it should be optimal for the fashion designer to be the principal and monitor/direct the sewer. In this example, the principal is also a labor input owner, i.e. he is an inside principal.
We accept Alchian and Demsetz’s basic arguments. However, there are important differences between their paper and our paper. First, for Alchian–Demsetz, it does not matter who becomes the monitor since the members are implicitly assumed to have identical initial advantages both in monitoring and in production. In contrast, we emphasize that the team members are heterogeneous in production advantages and in effectiveness of monitoring. It is this heterogeneity that dominates the assignment of principalship. Second, in Alchian–Demsetz, the monitor is specialized in monitoring, and hence he may not contribute production effort directly and becomes an outsider. In our paper, monitoring is only one of the functions of a particular team member who also makes direct production effort to the output. In other words, the monitor is still an insider of the team rather than a professional monitor as in their paper.2

We adopt an approach similar to that of Grossman and Hart (1986) in spirit. The difference between their approach and our approach is that they analyze residual rights of control over assets, while we analyze residual rights of control over actions. We confirm their conjecture (Grossman and Hart, 1986, p. 717) that their analyses of control rights over assets can be applied to the analyses of control rights over actions.

1.2. Introduction to the model

In our model, there are two members in a team.3 Each member’s efforts may consist of two types: production effort and monitoring effort, and he may have absolute or comparative advantages in certain activities. They both contribute labor inputs and jointly produce a product that they will share in a contract. We assume that the total output can be easily verified by a third party,4 but it is impossible to directly measure individual contributions to the joint output. Thus (individual), output monitoring is too costly, and input monitoring is adopted. There are four possible scenarios: (1) each member contributes only productive effort and there is no monitoring; (2) member 1 monitors member 2; (3) member 2 monitors member 1 and (4) they monitor each other. Zhang (1994) argues that (at least in the risk-neutral case) mutual-monitoring is Pareto-dominated by one-sided monitoring. The intuition is that mutual-monitoring incurs two monitoring costs, while one-sided monitoring incurs only one monitoring cost. Therefore, we will ignore case (4) and focus on the first three cases. We assume that production effort cannot be verified and enforced by a third party, but it can be induced by monitoring, directing and rewarding by a second party.

2 Budget-breaking mechanisms such as group penalties in Holmstrom (1982) may potentially solve the free-riding problem in teams. But as pointed out by Holmstrom, there are several limitations. First, such mechanisms are generally discontinuous. Second, if there is no independent third party (outsider), such mechanisms are not renegotiation-proof and hence not credible. Third, in the case of uncertainty, such mechanisms may become unfeasible due to endowment constraints. Fourth, it is not clear how to monitor or motivate the budget breaker such that he will not be opportunistic. Finally, as pointed out by Arrow (1985), such mechanisms suffer from multiple-equilibrium problem, and hence do not enforce the optimal outcome, although they permit it. In contrast, our proposed principalship mechanism is continuous, renegotiation-proof and feasible.

3 One member can be thought as a representative of a group who involves strategic or marketing activities, and the other may be a representative of a group who involves operational or producing activities.

4 For discussions of unobservable/unverifiable output, see Malcomson (1984). In such settings, a principal may have private information about the productivity of an agent. Thus an outsider principal can have moral hazard.
When there are no monitors/directors, the team members play a standard non-cooperative game. Since they cannot sign contracts on efforts, they face a ‘prisoners’ dilemma’ problem and the Nash equilibrium is inefficient. The constrained optimal incentive scheme becomes a residual sharing. In other words, it is not optimal to assign all residual to either party. The intuition is that fully assigning residual to either member will destroy the other’s incentive, and teamwork or complementarity makes the residual-holder’s effort useless. We find that the constrained optimal residual sharing of a party is strictly positively correlated to his relative importance in production, and negatively correlated to the degree of teamwork (complementarity) when he is more important than the other party in production. When one party is more important than the other in production, the other party forms an effort-bottleneck in production; in order to achieve an optimal result, higher complementarity of joint inputs requires the increase of the other party’s incentive to mitigate the bottleneck effect.

When one party monitors the other, there are two problems: effectiveness of monitoring/directing and limitations of authority. As indicated, we assume output monitoring is impossible and focus on input monitoring. In fact, when it is easy to directly meter individual output (i.e. marginal productivity), market exchange will be efficient in organizing economic activities as noted by Alchian and Demsetz (1972) (p. 778). Thus, input monitoring (e.g. checking that rules are followed) is central to the efficiency of an organization which supports team production. Besides on-site directing, watching and checking job design is a powerful monitoring/directing device and may take several forms: specific tasks and general rules which specify what procedures should be followed such as working hours and location, and also specify what activities one should not be involved in such as working in other firms. Relatively speaking, implementation of general rules incurs a higher fixed cost of monitoring/directing, while implementation of specific tasks incurs a more variable cost of monitoring/directing. To save the cost of monitoring, the principal may use random watching and random checks. In addition, there may exist economies of scale in monitoring; when one monitors multiple agents, the monitoring/directing cost per agent may decrease.

The above discussions on monitoring/directing can be summarized into a monitoring technology that specifies the relation between the principal’s monitoring/directing effort and the agent’s monitoring-induced labor input. In general, such relation is complex and the monitor does not have complete information of it. Our knowledge is very limited on monitoring technology (see Putterman and Skillman (1988) for more details). Since we shall focus on optimal assignment of principalship rather than on optimal design of monitoring schemes, we stress the fact that the principal’s monitoring/directing effort and the agent’s monitoring-induced effort are positively correlated, and assume that the monitor has perfect information of the monitoring technology.

The authority of monitoring and directing is limited by the agent’s right to withdraw. In other words, the package of a compensation scheme (‘carrots’) along with a monitoring scheme (‘sticks’) must be incentive compatible since the agent’s effort can only be induced by a second party, but not enforced by a third party. To induce the agent to make effort more than his self-incentive compatible effort under no monitoring, the principal needs to compensate the agent enough to offset his increased disutility. In other words, the purchasing price of principalship should make the agent no worse than his reservation utility under no monitoring.
We will show that both the principal’s production effort and monitoring effort increase as his residual share goes up. It is optimal to let the principal hold all the residual. That is, residual claims should match residual control as Alchian and Demsetz (1972) conjectured. Furthermore, when the monitoring cost converges to zero, the optimal solution with monitoring converges to the first best allocation as expected.

We now turn to the question of who purchases the right to monitor/direct and becomes the principal. We find that, as expected, if the two parties have the same advantage in production, it is optimal for the party with an advantage in monitoring to be the principal. Clearly, this arrangement saves monitoring cost. On the other hand, a somewhat less straightforward finding is that if the two parties have the same advantage in monitoring, then it is optimal for the party with an advantage in production to be the principal. The reasons are, first, this party has a larger marginal contribution in production, and full residual claims will induce him to input more efforts and hence more output and second, by assigning principalship to him, he is self-monitoring and this saves more monitoring cost. In the extreme case where one party is the sole production input contributor and the other party is an outsider, it is optimal for the insider to be the principal.

Therefore, when one member has (absolute) advantages in both production and monitoring, it is optimal to assign principalship to him regardless of the degree of teamwork. We conjecture that most real world cases fall into this category. But situations exist where one party has an advantage in production and the other in monitoring. 5 For instance, consider a coach and a figure skater. The coach has advantages in monitoring, while the skater is more important in ‘production’. In such cases, the degree of teamwork matters. A less straightforward interesting finding is that a higher degree of teamwork tends to call for the party with advantages in monitoring, but disadvantages in production, to be the principal. The reason is that with a higher degree of teamwork, output is more sensitive to the production effort of the less important member and less sensitive to that of the more important member. Therefore, the first member’s monitoring is more favored since that arrangement incurs less diminution of both members’ production incentives. 6

Finally, we show that principalship Pareto dominates partnership when at least one member is somewhat important in production and sufficiently effective in monitoring. On the other hand, when all members are ineffective in monitoring or when one is ineffective in monitoring and the other is somewhat effective in monitoring but not important in production, partnership is desirable. The former occurs when both members are involved in marketing activities (worker-in-the-dark) in industries such as law, accountancy and consulting.

We will present the general model in Section 2. Section 3 will focus on the class of CES production functions, and Section 4 concludes this paper.

2. The general model

Consider a team of two members: persons 1 and 2. They jointly produce a product by using the following technology

---

5 In such situations, each member has only comparative advantages.

6 This is similar to the earlier discussion on bottleneck effect for the case without monitoring.
where \( Y \) is output, \( \epsilon \) a random variable with mean zero, \( y \) the expected output, and \( a_1 \) and \( a_2 \) labor inputs of persons 1 and 2, respectively.\(^7\) Assume that the production function is increasing, concave and twice continuously differentiable, \( \partial^2 f / \partial a_1 \partial a_2 \geq 0 \) for all \( a_1, a_2 \geq 0 \) and \( f(0, 0) = 0 \). That is, we assume that \( f \) is supermodular. Given that \( f(0, 0) = 0 \), and \( f \) is supermodular, we have for all \( a_1, a_2 \geq 0 \), \( f(a_1, a_2) \geq f(a_1, 0) + f(0, a_2) \),\(^8\) which states that the team’s output is larger than or equal to the sum of individual outputs.

Next, assume that person \( i \)’s utility function is

\[
U_i(r_i, a_i, b_i) = r_i - C_i(a_i, b_i)
\]  

where \( r_i \) is income, \( a_i \) production effort, \( b_i \) monitoring effort, and \( C_i \) disutility function for \( i \in \{1, 2\} \). Thus the two members are risk-neutral. Assume that \( C_i \) is increasing, strictly convex and twice continuously differentiable with \( C_i(0, 0) = 0 \) for \( i \in \{1, 2\} \).

We may view production effort as transformation cost and monitoring effort as transaction cost. Monitoring effort is a form of interpersonal activity, and it is valuable only if there is moral hazard. In our setting, one person’s production effort can be ‘monitored’ by another person’s monitoring effort, but a person’s monitoring effort cannot be ‘monitored’ by another person’s monitoring effort. That is, a person’s monitoring effort can only be self-monitored. For simplicity, assume that the monitoring technology is as follows\(^9\)

\[
b_j \geq \frac{e_i}{\rho_j}
\]  

for \( i \neq j, \rho_j \geq 0 \), where \( b_j \) is person \( j \)’s monitoring/directing effort and \( e_i \) person \( i \)’s monitoring-induced production effort. One interpretation of the monitoring technology is that \( b_j \) is directing effort, including guiding, checking and correcting in the process and job design; it takes at least \( e_i / \rho_j \) effort for person \( j \) to direct person \( i \) to work \( e_i \). Another interpretation is that it takes at least \( e_i / \rho_j \) monitoring effort for person \( j \) to fully detect person \( i \)’s production effort \( e_i \). A higher \( \rho_j \) implies more effectiveness of \( j \)’s monitoring over \( i \). In particular, \( \rho_j = 0 \) implies that monitoring is technically impossible; \( \rho_j \to \infty \) implies that monitoring is perfect. In any case, in order for person \( i \) to accept person \( j \)’s directing and monitoring, person \( j \) needs to compensate person \( i \) enough to offset his increased disutility. Thus, monitoring is costly both directly (monitoring effort) and indirectly (disutility compensation).

\(^7\) In such a setting, a person is not able to infer the other person’s input based on the observation of the output and his own input.

\(^8\) See Milgrom and Roberts (1990, Theorem 2; Topkis (1978).

\(^9\) Such ‘proportional relation’ is the simplest formulation of monitoring technology. It seems that our main results (except specific relations such as inequality (Eq. (12)) in the next section) can be obtained under more general formulations of monitoring technology, although the explicit analyses become much less tractable. A linear monitoring technology can be specified as \( b_j \geq e_i / \rho_j + b_0 \), \( b_0 > 0 \), which may incorporate the costs of job design. For instance, we may consider fixed cost \( b_0 \) as the cost of designing and implementing general rules, and variable cost \( e_i / \rho_j \) as the cost of designing and implementing specific tasks in job design.
Finally, the reward contract (without monitoring) is a linear sharing scheme.  

\[ r_i = w_1 + \beta_i(Y - w_1 - w_2) \]  

(4)

for \( i = 1, 2, \) where \( w_1, w_2 \geq 0 \) are fixed terms, \( \beta_1 + \beta_2 \geq 0 \) are residual terms, and \( \beta_1 + \beta_2 = 1. \) Since we assume that the realization of (total) output can be verified by a third party, \( w_1, w_2, \beta_1, \beta_2 \) are enforceable by a third party. Assume further that expressions in Eqs. (1)–(4) are common knowledge.

Due to risk-neutrality, we only need to consider the expected output from now on. The first best arrangement is the solution to the following first order conditions.

\[ \frac{\partial f}{\partial a_1} (a_1, a_2) = \frac{\partial C_1}{\partial a_1} (a_1, 0), \quad \frac{\partial f}{\partial a_2} (a_1, a_2) = \frac{\partial C_2}{\partial a_2} (a_2, 0) \]  

(5)

If the two parties can sign (complete) contracts on their efforts, then the first best allocation can be achieved. We assume, however, the efforts are not contractible. This is the case when either efforts are not verifiable by a third party or it is too costly to specify all the details in a contract. The first best generally cannot be achieved when the efforts are not contractible except under perfect monitoring as will be seen later.

Consider the case without monitoring first. In this case, the two partners input efforts independently, and a Nash equilibrium will result. For a given \( i, \) person \( i \)'s reaction function is the solution of the first order condition.

\[ \beta_i \frac{\partial f}{\partial a_i} (a_1, a_2) = \frac{\partial C_i}{\partial a_i} (a, 0) \]  

(6)

Our assumptions on \( f \) and \( C \) imply

\[ \frac{\partial}{\partial a_i} \left( \beta_i \frac{\partial f}{\partial a_1} (a_1, a_2) - \frac{\partial C_i}{\partial a_i} (a_i, 0) \right) \neq 0 \]

Hence, by the implicit function theorem, there exists a continuous reaction function defined in Eq. (6) for each \( i. \) The intersection of these two reaction curves determines the Nash equilibrium. Assume there exists an unique Nash equilibrium without monitoring. Then by the implicit function theorem, the effort functions \( a_1(\beta_1) \) are \( a_2(\beta_1) \) continuous in \( \beta_1. \)

Let \( \pi(\beta_1) \) be the total welfare at share \( \beta_1 \) without monitoring, i.e. \( \pi(\beta_1) = f(a_1(\beta_1), a_2(\beta_1)) - C_1(a_1(\beta_1), 0) - C_2(a_2(\beta_1), 0). \) Then \( \pi \) is a continuous function. By Weierstrass theorem, \( \pi \) achieves its maximum on \((0, 1). \) Let \( \beta_1^\# \) be the optimal sharing at which the total welfare is maximized. Clearly, we have \( 0 < \beta_1^\# < 1, \) i.e. the (constrained) optimal distribution scheme is a residual sharing and it is not optimal to assign all residual to either party.

It is not surprising that the maximum welfare attained at Nash equilibrium without monitoring is less than the first best welfare. This is because the Nash equilibrium efforts without

---

10 For the logic of linear compensation formulas, see Holmstrom and Milgrom (1987).

11 Technically we assume that \( w_1 + w_2 \leq Y \) such that the fixed terms are riskless, where \( Y \) denotes the lowest possible output.

12 This is the case when monitoring/directing is either technically impossible or economically inferior to no monitoring. See more discussions later.

13 If there are several maximizers, choose any one.
monitoring are smaller than the first best efforts. That is, there is under-production due to the free-riding problem (also see Holmstrom (1982), pp. 326–327).

We now turn to the case with monitoring. Suppose that person 1 monitors person 2. Person 1 will purchase the rights to monitor and direct person 2. Such rights are limited residual rights of control over actions. They are limited by the right to withdraw and the hidden action of person 2. Person 1 can use both ‘sticks’ and ‘carrots’ to induce person 2’s incentive. If we view monitoring and directing as ‘sticks’, then the purchasing price can be viewed as ‘carrots’. Given person 2’s self-effort \( a_2 \), person 1’s problem is as follows:

\[
\max w_1 + \beta_1(f(a_1, e_2) - w_1 - w_2) - F - C_1(a_1, b_1), \quad a_1, b_1, e_2, F \geq 0 \quad (7)
\]

s.t.

\[
b_1 \geq \frac{e_2}{\rho_1}, \quad F + w_2 + \beta_2(f(a_1, e_2) - w_1 - w_2) - C_2(\rho_1 b_1, 0)
\geq w_2 + \beta_2(f(a_1, a_2) - w_1 - w_2) - C_2(a_2, 0)
\]

where \( F \) is the price that person 1 pays to purchase the rights to monitor and direct person 2. The LHS of the above constraint is what person 2 can get if he accepts being monitored and directed; person 2 accepts 1’s directing and inputs \( e_2 \), and receives disutility compensation \( F \). The RHS is what person 2 can get if he does not accept being monitored and directed; person 2 chooses his own input and does not receive disutility compensation. The RHS forms the bargaining status quo of person 2. The constraint says that person 2’s utility under monitoring by person 1 must be no less than that under no monitoring. In equilibrium, they are equal (and \( b_1 = e_2/\rho_1 \)). When \( e_2 \) goes up, person 2’s disutility increases, and \( F \) generally goes up. That is, to induce person 2 to accept person 1’s ‘more directing and monitoring’, person 1 needs to compensate person 2 more to offset his increased disutility. For monitoring to occur in equilibrium, it must be the case \( \rho_1 b_1 > a_2 \), i.e. the monitor-induced incentive is larger than the self-incentive of person 2 under no monitoring.

The above scheme can be enforced by different mechanisms. The fixed payments \( w_1, w_2 \) and the residual shares \( \beta_1, \beta_2 \) are enforceable by a third party. The price of purchasing principalship \( F \) and the monitor-induced effort \( e_2 \) are enforceable by second parties (i.e. mutual enforcing). The self-efforts \( a_1, b_1, a_2 \) are enforceable by first parties (i.e. self-incentive driven). Incentive compatibility conditions assure first-party enforcement. The reader can also see Holmstrom (1996, p. 9) for discussions of the enforcement of promised payment by the monitor to other hired inputs.

The solution to Eq. (7) gives person 1’s reaction functions \( a_1 \) and \( b_1 \) to person 2’s self-effort \( a_2 \). These two reaction functions along with person 2’s reaction function defined in Eq. (6) \((i = 2)\) determine the Nash equilibrium with monitoring. The first order conditions for Eq. (7) are as follows:\(^\text{14}\)

\[
\begin{align*}
\frac{\partial f}{\partial a_1}(a_1, \rho_1 b_1) - \beta_2 \frac{\partial f}{\partial a_2}(a_1, a_2) &= \frac{\partial C_1}{\partial a_1}(a_1, b_1), \\
\frac{\partial f}{\partial a_2}(a_1, \rho_1 b_1) &= \frac{1}{\rho_1} \frac{\partial C_1}{\partial b_1}(a_1, b_1) + \frac{\partial C_2}{\partial a_2}(\rho_1 b_1, 0)
\end{align*}
\]

\(^{14}\)In the notation of the following partial derivatives, \( a_1 \) and \( b_1 \) serve as generic variables for arguments. Note that the values of arguments could be different from the generic variables.
Assume that there exists an unique Nash equilibrium with monitoring. Then by the implicit function theorem, the effort functions are continuous in the parameters of sharing and monitoring. Let \( \Pi_1(\rho_1, \beta_1) \) be the welfare function when person 1 monitors person 2, i.e. \( \Pi_1 \equiv f(a_1, \rho_1b_1) - C_1(a_1, b_1) - C_2(\rho_1b_1, 0) \), where \( a_1 \) and \( b_1 \) are the solutions to Eqs. (6) and (8) for \( i = 2 \) and hence are functions of \( \beta_1 \) and \( \rho_1 \). Then \( \Pi_1(\rho_1, \beta_1) \) is a continuous function. By using Eq. (8), it can be shown that

\[
\frac{\partial \Pi_1}{\partial \beta_1} = \beta_1^2 \frac{\partial f}{\partial a_1}(a_1, a_2) \frac{\partial a_1}{\partial \beta_1}, \quad \frac{\partial \Pi_1}{\partial \rho_1} = \beta_1^2 \frac{\partial f}{\partial a_1}(a_1, a_2) \frac{\partial a_1}{\partial \rho_1} + b_1 \frac{\partial C_1}{\partial b_1}(a_1, b_1) \tag{9}
\]

For the class of CES production functions, we will show that person 1’s production effort increases with both his effectiveness of monitoring and residual share. Then by Eq. (9), welfare also increases with both person 1’s effectiveness of monitoring and residual share. Thus, it is optimal to let the principal hold all the residual. Residual claims should match welfare also increases with both person 1’s effectiveness of monitoring and residual share. Then by Eq. (9), welfare also increases with both person 1’s effectiveness of monitoring and residual share. Then by Eq. (9), welfare also increases with both person 1’s effectiveness of monitoring and residual share. Thus, it is optimal to let the principal hold all the residual. Residual claims should match welfare also increases with both person 1’s effectiveness of monitoring and residual share. Then by Eq. (9), welfare also increases with both person 1’s effectiveness of monitoring and residual share. Thus, it is optimal to let the principal hold all the residual. Residual claims should match welfare also increases with both person 1’s effectiveness of monitoring and residual share. Then by Eq. (9), welfare also increases with both person 1’s effectiveness of monitoring and residual share. Thus, it is optimal to let the principal hold all the residual. Residual claims should match welfare also increases with both person 1’s effectiveness of monitoring and residual share.

It is easy to see that \( \Pi_1(0, 1) < \pi(\beta_1^0) < \Pi_1(\infty, 1) \). By the intermediate value theorem, there exists \( \rho_1^* \in (0, \infty) \) such that \( \Pi_1(\rho_1^*, 1) = \pi(\beta_1^0) \). When monitoring effectiveness is above this critical level, principalship Pareto dominates partnership; otherwise the cost of monitoring is higher than the gain resulting from monitoring, making partnership desirable.

Let \( \Pi_2(\rho_2, \beta_2) \) be the welfare function when person 2 monitors person 1. When \( \Pi_1(\rho_1, 1) > \Pi_2(\rho_2, 1) \), it is optimal for person 1 to monitor person 2 provided principalship Pareto dominates partnership. Assume that for any pre-assigned principalship, the two parties can costlessly bargain for re-assignment of principalship (through changing \( w_1, w_2 \)). Then (by ‘Coase theorem’) their bargaining will always result in an efficient assignment of principalship since the two parties have symmetric information about bargaining in our setting.

3. Optimal principalship for the class of CES production functions

In order to derive more specific results, we examine the class of CES production functions in this section. That is, the production function \( f \) in Eq. (1) becomes

\[
f(a_1, a_2) = (\alpha_1a_1^{1-\gamma} + \alpha_2a_2^{1-\gamma})^{1/(1-\gamma)} \tag{10}
\]

where \( \alpha_1, \alpha_2, \gamma \in [0, 1] \) and \( \alpha_1 + \alpha_2 = 1 \). It is easy to see that \( f \) is increasing, concave and twice continuously differentiable, \( \partial^2 f/\partial a_1 \partial a_2 \geq 0 \) for all \( a_1, a_2 \geq 0 \) and \( f(0, 0) = 0 \).

In Eq. (10), \( \alpha_i \) is person \( i \)’s effort elasticity of output. In the context of this paper, we take \( \alpha_i \) as a measure of person \( i \)’s importance or advantage in production. When \( \alpha_i > (\gamma)1/2 \), person \( i \) is more (less) important than person \( j \). When \( \alpha_1 = \alpha_2 = 1/2 \), the two members

\[15\] Note that our proposed principalship mechanism is continuous, renegotiation-proof and feasible.
are equally important. Parameter $\gamma$ is the ‘elasticity of complementarity’ of the two efforts in production. Note $\frac{\partial^2 f}{\partial a_1 \partial a_2}$ increases with $\gamma$. That is, the larger $\gamma$ is, the more one member’s effort will increase the other member’s marginal productivity. When $\gamma = 0$, Eq. (10) reduces to a linear function. In this case, $f$ is an additive function and teamwork does not increase marginal productivity. When $\gamma = 1$, Eq. (10) becomes a Cobb–Douglas function and teamwork increases marginal productivity. For this reason, we define $\gamma$ as ‘the degree of teamwork’. $\gamma = 0$ implies no teamwork and $\gamma = 1$ implies pure teamwork.\footnote{Strictly speaking, Leontief technology ($\gamma = 1$) represents pure teamwork. Since the case of $\gamma = \infty$ is not tractable, we restrict ourselves to $0 \leq \gamma \leq 1$.}

To simplify our analyses further, assume that the disutility function $C_i$ in Eq. (2) takes the following quadratic form.

$$C_i(a_i, b_i) = \frac{1}{2}a_i^2 + \frac{1}{2}b_i^2$$

for $i = 1, 2$. Clearly, $C_i$ is increasing, strictly convex and twice continuously differentiable, and $C_i(0, 0) = 0$ for $i = 1, 2$. Finally, monitoring technology is the same as Eq. (3).

We now present the propositions and corollaries and all the proofs are given in Appendix A. Consider the case without monitoring first. We have

**Proposition 1.** In partnership, (i) the optimal share $\beta_i^T$ strictly increases with person $i$’s effort elasticity $\alpha_i$; (ii) the optimal share $\beta_i^T$ increases with the degree of teamwork $\gamma$ for $\alpha_i < 1/2$ and decreases with $\gamma$ for $\alpha_i < 1/2$.

**Corollary 1.** For $\alpha_i \neq 0, 1$, we have $0 < \beta_i^T < 1$.

Proposition 1 can be demonstrated in Fig. 1 below. Note that as $\gamma$ increases, $\beta_i^T$ is getting closer to $1/2$, i.e. a higher degree of teamwork tends to result in a more equal residual share.

Proposition 1 and Corollary 1 tell us that it cannot be optimal to fully assign the ‘principalship’ to either of the two members when monitoring is technically impossible. The optimal assignment requires a balance of incentives between the two members, and the degree of teamwork matters. In particular, the optimal residual share held by each member should be positively related to his relative importance in production as common sense suggests, and negatively correlated to the degree of team work when he is more important than the other party in production. The more important his production effort is, the bigger his residual share. However, in general this relationship is not linear, and it is rather complex. There are two effects at work in determining the deviation of the optimal residual share from the relative importance in production.

The first is the output effect. From the point of view of maximization of the total output, the residual share assigned to the more important member should be more than proportional to his relative importance, i.e. $\beta_i^{OT} > \alpha_i$ for $\alpha_i > 1/2$ and $\beta_i^{OT} < \alpha_i$ for $\alpha_i < 1/2$ (here we define $\beta_i^{OT}$ as the residual share which maximizes the output), unless $\gamma = 1$ in which case $\beta_i^{OT} \equiv \alpha_i$ for all $0 \leq \alpha_i \leq 1$ (see the Cobb–Douglas case in Appendix B). The reason is that at $\beta_i = \alpha_i$ the marginal productivity of the more important member is greater than the marginal productivity of the less important member; a small shift to favor the more important
member will induce him to work more, which will more than offset the disincentive of the less important member, as long as the production function is not pure teamwork.

The second is the cost effect. Under the assumption of identical preferences, at $\beta_i = \alpha_i$, the more important member incurs higher marginal cost (in terms of disutility) than the less important member in providing effort. Therefore, it is desirable to deviate from $\beta_i = \alpha_i$ from the point of view of the cost reduction. Whether the optimal residual share $\beta_i^*$ is less than, equal to or greater than the relative importance depends on the dominance of one effect over the other, which in turn depends on the degree of teamwork through the interdependence of marginal productivities. A higher degree of teamwork implies higher interdependence and hence a lower productivity advantage of the more important member at $\beta_i = \alpha_i$. In particular, when the production function is pure teamwork ($\gamma = 1$) the two marginal productivities are equalized at $\beta_i = \alpha_i$ and therefore the output effect disappears at $\beta_i = \alpha_i$, and the cost effect implies that the less important member should be assigned a residual share more than proportional to his relative importance (in other words, the more important member should be assigned a residual share less than proportional to his relative importance).\(^{17}\)

Our analyses can be applied to the senior–junior partner relationship. In a partnership firm, senior partners are more important than junior partners, and the former also share more output than the latter. We now know this arrangement is in fact consistent with efficiency.

\(^{17}\) The above discussion also suggests that in a general form of cost (disutility) functions, the more costly one’s effort is, the smaller his residual share is.
Our theory further predicts that when the partners have a greater dependence on each other, the senior partners need to give up more shares to junior partners to mitigate the bottleneck effects. This theoretical prediction is confirmed by the empirical evidence in law firms (Gilson and Mnookin, 1985). They find that compensation differs to a less extent than contribution to the firm as measured by observables.

Next, suppose person 1 monitors person 2, we have our central results in the following proposition and corollary.

**Proposition 2.** In the case where person 1 monitors person 2, we have (i) person 1’s production effort $a_1$ increases with both his residual share $\beta_1$ and his effectiveness of monitoring $\rho_1$; (ii) person 1’s monitoring effort $b_1$ increases with his residual share $\beta_1$; and (iii) total welfare increases with both person 1’s residual share $\beta_1$ and his effectiveness of monitoring $\rho_1$.

**Corollary 2.** In the case where person 1 monitors person 2, the optimal residual share is $\beta_1 = 1$.

Proposition 2 tells us that both the principal’s production effort and monitoring effort are positively correlated to his residual share. Corollary 2 says that it is optimal to let the principal hold all the residual. Thus we have confirmed the celebrated conjecture by Alchian and Demsetz (1972), that residual claims should match residual control. This is an intuitive but important result. Furthermore, as noted in the general case in the last section, when monitoring becomes sufficiently effective, full principalship can approximate the first best allocation. In this sense, we view principalship as the primary incentive mechanism in teams.

The intuition behind Proposition 2 and Corollary 2 is that when person 1 monitors person 2, it is person 1 who effectively determines all production inputs (including his own production effort and person 2’s production effort) subject to purchasing principalship from person 2. Increasing person 1’s share of output will decrease person 2’s bargaining status and hence lower the ‘purchasing price’ of principalship. Thus, the marginal benefits of both person 1’s production effort and monitoring effort increase with his share of output. When person 1’s residual share of output increases to 1, the free-riding problem disappears (up to monitoring costs).

The above result provides an explanation to the pervasive phenomenon of asymmetric contractual arrangements between different members of a team. For instance, in a classical (private) firm, the entrepreneur claims the full residual and directs/monitors workers, and workers are entitled to take fixed wages and are obliged to obey the authority of the entrepreneur within certain limits. Our theory suggests such a ‘corner solution’ is in fact an efficient arrangement since full residual claims induce self-monitoring of the monitor.

---

18 The relationship between $b_i$ and $\rho_i$ is not monotone. Typically monitoring effort first increases and then decreases with effectiveness of monitoring. This is because monitoring itself is not productive and its value comes only from affecting the agent’s incentive. An observation is that an employer equipped with a closed circuit television surveillance spends less time in monitoring than otherwise.

19 This result implies that the agent takes fixed payment and hence the result also holds if the agent is risk-averse.
We now identify the determinants of who should be the monitor. We have\textsuperscript{20}

**Proposition 3.** For $0 \leq \gamma < 1$, $\Pi_1(\rho_1, 1) \geq \Pi_2(\rho_2, 1)$ if and only if

$$
\left(1 - \left(\frac{\rho_1^2}{1 + \rho_1^2}\right)^{(1-\gamma)/(1+\gamma)}\right)^{\frac{1}{2}} \leq \left(\frac{\alpha_1}{\alpha_2}\right)^{2/(1+\gamma)} \left(1 - \left(\frac{\rho_2^2}{1 + \rho_2^2}\right)^{(1-\gamma)/(1+\gamma)}\right) 
$$

\text{(12)}

**Corollary 3.** For $0 \leq \gamma < 1$: (i) if $\alpha_1 \geq \alpha_2$ and $\rho_1 \geq \rho_2$, then $\Pi_1(\rho_1, 1) \geq \Pi_2(\rho_2, 1)$; (ii) if $\alpha_i < \alpha_j$ and $\rho_i > \rho_j$, then $\Pi_i$ more likely dominates $\Pi_j$ as $\gamma$ increases for $i \neq j$.

The key point of Proposition 3 is that the optimal assignment of principalship depends on the interaction between relative importance in production, monitoring technology and the degree of teamwork. Corollary 3 implies that if the two parties have the same advantage in production, it is optimal for the party with an advantage in monitoring to be the principal. Clearly, this arrangement saves monitoring cost. A somewhat less intuitive finding is that if the two parties have the same advantage in monitoring, it is optimal for the party with an advantage in production to be the principal. This is because this party has a larger marginal contribution in production, and full residual claim will induce him to input more efforts and hence more output. In addition, by assigning principalship to him, he is self-monitoring and this saves more monitoring cost. Therefore, it is optimal to assign principalship to the party with (absolute) advantages in both production and monitoring.

However, one’s absolute advantages in both production and monitoring are a sufficient, but not a necessary, condition for his principalship to Pareto dominate the other’s. The optimal assignment may require a less important member be the principal if he enjoys a big enough advantage in monitoring, or a less effective monitor be the principal if his role is dominant enough in production. Furthermore, Corollary 3 (ii) says that whenever two members have different advantages in production and monitoring, the degree of teamwork matters. An interesting and nontrivial finding is that an increase in the degree of teamwork tends to strengthen the role of monitoring advantage but weaken that of production advantage in determining the optimal assignment of principalship. It is more likely to be optimal to let the less important member monitor the more important member when the degree of teamwork is high than when it is low, given that the first member has an advantage in monitoring. The intuition is that with a higher degree of teamwork, output is more sensitive to the production effort of the member who is less important in production but more effective in monitoring, and less sensitive to the production effort of the member who is the reverse. The first member’s monitoring is more favorable because it incurs less diminution of both members’ production efforts.\textsuperscript{21}

\textsuperscript{20} Recall $\Pi_i(\rho_i, \gamma)$ is the welfare function when $i$ monitors $j$, and $\alpha_i$ and $\gamma$ are suppressed in the notation.

\textsuperscript{21} This finding can shed some light on the phenomenon that the managers/monitor in a manufacturing firm is often not, or not the best, expert of production activities, but he or she is the best expert of monitoring activities. He or she may know little about engineering, but he or she knows very well who are good engineers in what areas.
The last proposition compares principalship with partnership (recall $\pi(\beta^p_i)$ is the welfare function for partnership valued at the optimal share).

**Proposition 4.** (i) For $i = 1, 2$, there exists $\rho^*_i \in (0, \infty)$ such that $\Pi_i(\rho^*_i, 1) = \pi(\beta^p_i)$; (ii) if $\rho_1 < \rho^*_1$ and $\rho_2 < \rho^*_2$, then $\pi(\beta^p_i) > \Pi_i(\rho_i, 1)$ for $i = 1, 2$; (iii) if for some $\rho_i > \rho^*_i$, then $\Pi_i(\rho_i, 1) > \pi(\beta^p_i)$.

Note $\rho^*_i$ is the ‘minimum sufficient effectiveness’ of monitoring for member $i$’s principalship to Pareto dominate partnership. Proposition 4 says that when the monitoring effectiveness of both members is lower than their minimum sufficient level, partnership Pareto dominates principalship. On the other hand, as long as at least one member is sufficiently effective in monitoring, it is optimal to adopt principalship instead of partnership. Intuitively, the principal’s efforts are motivated by self-interest, while the agent’s effort can only be induced by the principal’s monitoring/directing along with disutility compensation. Under partnership, both members have ‘half’ self-incentive to work; under principalship, the principal gains the full self-incentive while the agent loses all the self-incentive (or gains the full incentive to shirk). Principalship is preferred if and only if the principal’s monitoring is so effective that the loss of the agent’s self-incentive can be more than offset by the monitoring-induced incentive. Given that principalship is desirable, it is optimal for person $i$ to monitor person $j$ whenever $\Pi_i(\rho_i, 1) > \Pi_j(\rho_j, 1)$.

As discussed in the last section, bargaining will always result in an efficient assignment of principalship since the two parties have symmetric information about bargaining in our setting.

Fig. 2 demonstrates the relation between $\rho^*_i$ and $\alpha_i$ for $\gamma = 0$ and $\gamma = 1$. Note that the vertical axis is $\rho^*_i/1 + \rho^*_i$ instead of $\rho^*_i$ (and therefore maximum 1 represents perfect monitoring $\rho^*_i = \infty$). The figure suggests that the minimum sufficient effectiveness of monitoring $\rho^*_i$ decreases with $\alpha_i$ for a given degree of teamwork. It follows that principalship

![Fig. 2. The relationship between $\rho^*_i$ and $\alpha_i$.](image-url)
is more likely to be preferred when the production advantage and the monitoring advantage are held by the same member than when they are held separately. 22

4. Conclusions

Instead of focusing on how a principal designs an optimal incentive scheme for agents as in the agency literature, we have shifted attention to determinants of principalship. As a first step, we have investigated the optimal assignment of principalship in teams. We emphasize that when it is costly to sign contracts on efforts, it may be optimal to let one party to purchase the rights to monitor and to direct, and claim full residual. Principalship is the purchase of these rights. The incentive structure changes when principalship shifts from one member to another or to partnership. Under partnership (residual sharing), both members have ‘half’ self-incentives. When one member becomes the principal and the other becomes the agent, the former gains full self-incentive but the latter has lost self-incentive and can only be induced to work by the former’s monitoring along with disutility compensation. The optimal assignment is optimally balancing this trade-off.

In particular, we have shown how the optimal assignment of principalship depends on the interaction between the relative importance of each member, the effectiveness of monitoring technology, and the degree of teamwork. We find that: (1) partnership is more likely to be preferred when monitoring is technically impossible or less effective and both team members are ‘somewhat’ important. Given partnership, a higher degree of teamwork tends to result in a more equal residual sharing scheme; (2) when one member monitors/directs the other, it is optimal to entitle the monitor/director to take full residual. This is an intuitive but important result; (3) principalship is more likely to be preferred when a single member possesses advantages in both production and monitoring. Given that principalship is to be adopted, it is optimal to let the member with both advantages be the principal regardless of the degree of teamwork. However, when the two advantages are held separately, the degree of teamwork matters. An interesting and nontrivial finding is that an increase in the degree of teamwork will favor the member with an advantage in monitoring but disfavor the member with an advantage in production.

Our paper can shed some light upon institutional arrangements of various organizations. In a typical firm, there are two types of insiders: the ‘marketing member’ of deciding what to do and how to do it, and the ‘producing member’ of implementing the decisions. In a classical capitalist firm, the marketing member is the principal who claims the residual and possesses monitoring/directing authority, while the producing member is the agent who receives a fixed payment and is monitored/directed by the marketing member. In conventional terminology, the former is called ‘the entrepreneur’ and the latter a ‘worker’. Our model

22 If we ‘interpret’ partnership as nonintegration and principalship as integration, then our results are ’similar’ to that of Grossman and Hart (1986) (pp. 716–717). They have shown that “... integration is optimal when one firm’s investment decision is particularly important relative to the other firm’s, whereas nonintegration is desirable when both investment decisions are ‘somewhat’ important’ (Grossman and Hart, 1986, pp. 716–717). We have shown that principalship is optimal when at least one member is important in production and effective in monitoring, whereas partnership is desirable when both members are ‘somewhat’ important in production and ineffective in monitoring.
legitimizes this positional transformation from ‘marketing member’ to ‘entrepreneur’ and from ‘producing member’ to ‘worker’. As Knight (1964) pointed out a long time ago (1921, p. 268), under uncertainty, “the actual execution of activity becomes in a real sense a secondary part of life; the primary problem or function is deciding what to do and how to do it”. This means that the marketing member is more important. The marketing member decides what to do and how to do it, and his activities are inventive and creative, while the producing member simply transforms inputs into output physically according to the marketing member’s decisions and his activities are almost routine. In other words, the marketing member is a ‘worker-in-the-dark’ and the producing member is a ‘worker-in-the-light’. Therefore it is more effective for the marketing member to monitor the producing member than otherwise. After all, a glance at the producing member will reveal whether he is working, while a stare at the marketing member may tell little about what he is doing. To sum up, the marketing member is not only the most important but also more difficult to be monitored. Therefore, it is optimal to let him monitor himself by assigning principalship to him. A worker-owned firm fails because principalship is assigned to the members who are less important and less effective in monitoring (see Yang and Ng (1995) for a model using a different approach but containing similar ideas).

Our model also predicts that partnership is more likely to prevail in a firm in which members are equally important and highly complementary in production, and equally difficult to monitor. In reality, partnership is commonly observed in firms with ‘artistic’ or ‘professional’ inputs (Alchian and Demsetz, 1972) (p. 786) such as lawyers, advertising specialists, doctors, accountants, consultants, academic researchers and teachers. In these industries, all members may involve ‘marketing’, and ‘producing’ is trivial;23 the output depends much more on the efficient uses of all members’ intelligence rather than on how many hours they spend in the office. In other words, all members are ‘workers-in-the-dark’. This makes monitoring very ineffective. As a result, a partnership contract provides higher overall incentives than a principalship does.24

Acknowledgements

We are particularly grateful to Donald Hay, Meg Meyer, James Mirrlees, and an anonymous referee for their helpful comments. Thanks also go to Christopher Bliss, Paul Klemperer, James Malcomson, Guofu Tan, John Vickers, Yijiang Wang and the audiences of the Royal Economics Society Annual Conference and the Seminars at Nuffield College and Institute of Economics and Statistics in Oxford University in 1994, and the Allied Social Science Annual Meeting at Chicago in 1998.

23 ‘Producing persons’ such as typists are normally not partners.
24 When should an economics professor treat her research assistant as a co-author or only acknowledge him in a footnote? An observation is that when the research requires assistance from brain, the research assistant appears as a co-author; on the other hand, when the research needs assistance mainly from hands (collecting and calculation of data), the research assistant is ‘gratefully acknowledged’. Co-authorship will always motivate the research assistant to work harder. But that might damage the professor’s incentive so seriously that papers are never finished.
Appendix A. The CES case

Part (i): The first-best solutions are

\[ a_{1}^{FB} = \alpha_1 \left[ \alpha_1 + \alpha_2 \left( \frac{\alpha_2}{\alpha_1} \right)^{(1-\gamma)/(1+\gamma)} \right]^{\gamma/(1-\gamma)} \]

\[ a_{2}^{FB} = \alpha_2 \left[ \alpha_2 + \alpha_1 \left( \frac{\alpha_1}{\alpha_2} \right)^{(1-\gamma)/(1+\gamma)} \right]^{\gamma/(1-\gamma)} \]

\[ y^{FB} = \left[ \alpha_1^{2/(1+\gamma)} + \alpha_2^{2/(1+\gamma)} \right]^{(1+\gamma)/(1-\gamma)} \]

\[ I_{i}^{FB} = \frac{1}{2} \left[ \alpha_1^{2/(1+\gamma)} + \alpha_2^{2/(1+\gamma)} \right]^{(1+\gamma)/(1-\gamma)} \]

Part (ii): When there are no monitors, member \( i \) chooses his effort \( a_i \) to maximize his expected utility, taking the other’s effort \( a_j \) as given

\[ \max_{a_i} \beta_i [\alpha_1 a_i^{1-\gamma} + \alpha_2 a_2^{1-\gamma}]^{1/(1-\gamma)} - \frac{1}{2} a_i^2, \quad i = 1, 2 \]

Solving the two maximization problems gives two reaction functions.

\[ a_1^{1+\gamma} = \beta_1 \alpha_1 \left[ \alpha_1 a_1^{1-\gamma} + \alpha_2 a_2^{1-\gamma} \right]^{1/(1-\gamma)} \]

\[ a_2^{1+\gamma} = \beta_2 \alpha_2 \left[ \alpha_1 a_1^{1-\gamma} + \alpha_2 a_2^{1-\gamma} \right]^{1/(1-\gamma)} \]

The Nash equilibrium is

\[ a_i^* = \beta_1 \alpha_1 \left[ \alpha_1 + \alpha_2 \left( \frac{\beta_1 \alpha_1}{\beta_2 \alpha_2} \right)^{(1-\gamma)/(1+\gamma)} \right]^{\gamma/(1-\gamma)} \]

\[ a_2^* = \beta_2 \alpha_2 \left[ \alpha_2 + \alpha_1 \left( \frac{\beta_1 \alpha_1}{\beta_2 \alpha_2} \right)^{(1-\gamma)/(1+\gamma)} \right]^{\gamma/(1-\gamma)} \]

Substituting the Nash equilibrium efforts into the production function and the welfare function, we have

\[ y = [\beta_1^{(1-\gamma)/(1+\gamma)} a_1^{2/(1+\gamma)} + \beta_2^{(1-\gamma)/(1+\gamma)} a_2^{2/(1+\gamma)}]^{(1+\gamma)/(1-\gamma)} \]

\[ \pi = \frac{1}{2} [\beta_1^{(1-\gamma)/(1+\gamma)} a_1^{2/(1+\gamma)} \beta_1^{(1-\gamma)/(1+\gamma)} a_2^{2/(1+\gamma)}]^{2\gamma/(1-\gamma)} \]

\[ \times [\beta_1^{(1-\gamma)/(1+\gamma)} (1 + \beta_2 \alpha_2^{2/(1+\gamma)}) \beta_2^{(1-\gamma)/(1+\gamma)} (1 + \beta_1 \alpha_1^{2/(1+\gamma)})] \]

Differentiating \( \pi \) with respect to \( \beta_i \) (note that \( \beta_1 + \beta_2 = 1 \) and \( \alpha_1 + \alpha_2 = 1 \)) and doing some tedious manipulations, we have the following first-order condition holds for the optimal residual share \( \beta_i^\pi \)
\[ A = (1 - \beta_i)^2 \left( \frac{\alpha_i}{1 - \alpha_i} \right)^{2/(1 + \gamma)} - \beta_i^2 \left( \frac{1 - \beta_i}{\beta_i} \right) \left( \frac{1 - \alpha_i}{\alpha_i} \right)^{2/(1 + \gamma)} \]

\[ - \left( \frac{2\beta_i - 1}{1 + \gamma} \right) \left( \frac{1 - \beta_i}{\beta_i} \right)^{(1 - \gamma)/(1 + \gamma)} = 0 \]

For the second-order condition, differentiating \( A(\beta_i, \alpha_i, \gamma) \) with respect to \( \beta_i \) and simplifying the derivative with the first-order condition successively, we have

\[
\frac{\partial A}{\partial \beta_i} = - \left( \frac{2\gamma}{(1 + \gamma)\beta_i} \right) \left( \frac{1 - \beta_i}{\beta_i} \right) \left( \frac{\alpha_i}{1 - \alpha_i} \right)^{2/(1 + \gamma)}
- \beta_i \left( \frac{1 - \beta_i}{\beta_i} \right)^{2(1 - \gamma)/(1 + \gamma)} \left( \frac{1 - \alpha_i}{\alpha_i} \right)^{2/(1 + \gamma)}
- \left( \frac{2\gamma}{1 + \gamma} \right) \left( \frac{1 - \beta_i}{\beta_i} \right)^{(1 - 3\gamma)/(1 + \gamma)} \left( \frac{1 - \alpha_i}{\alpha_i} \right)^{2/(1 + \gamma)}
- \frac{1}{(1 + \gamma)\beta_i} \left( \frac{1 - \beta_i}{\beta_i} \right)^{-2\gamma/(1 + \gamma)} < 0
\]

This means that the second-order condition is satisfied and the implicit function theorem applies.

**Proof of Proposition 1.** Part (i): By the implicit function theorem, we have \( \frac{\partial \alpha_i}{\partial \beta_i} = -\left( \frac{\partial A}{\partial \alpha_i} \right) \left( \frac{\partial A}{\partial \beta_i} \right) \). Since \( \frac{\partial A}{\partial \beta_i} < 0 \), the sign of \( \frac{\partial \alpha_i}{\partial \beta_i} \) is the same as the sign of \( \frac{\partial A}{\partial \alpha_i} \).

Differentiating \( A(\beta_i, \alpha_i, \gamma) \) with respect to \( \alpha_i \) and rearranging with the first-order condition successively, we have

\[
\frac{\partial A}{\partial \alpha_i} = [\alpha_i^{-1} + (1 - \alpha_i)^{-1}] \left( 1 - \beta_i \right)^2 \left( \frac{\alpha_i}{1 - \alpha_i} \right)^{2/(1 - \gamma)}
+ \beta_i^2 \left( \frac{1 - \beta_i}{\beta_i} \right)^{2(1 - \gamma)/(1 + \gamma)} \left( \frac{1 - \alpha_i}{\alpha_i} \right)^{2/(1 + \gamma)} > 0
\]

This proves \( \frac{\partial \alpha_i}{\partial \beta_i} > 0 \) for all \( \gamma \).

Part (ii): Differentiating \( A(\beta_i, \alpha_i, \gamma) \) with respect to \( \gamma \) and rearranging, we have

\[
\frac{\partial A}{\partial \gamma} = 2(1 - \beta_i)^2 \left( \frac{\alpha_i}{1 - \alpha_i} \right)^{2/(1 + \gamma)} \left[ \ln \left( \frac{(1 - \beta_i)(1 - \alpha_i)}{\beta_i \alpha_i} \right) + \frac{1 + \gamma}{2} \right]
+ 2\beta_i^2 \left( \frac{1 - \beta_i}{\beta_i} \right)^{2(1 - \gamma)/(1 + \gamma)} \left( \frac{1 - \alpha_i}{\alpha_i} \right)^{2/(1 + \gamma)}
\times \left[ \ln \left( \frac{(1 - \beta_i)(1 - \alpha_i)}{\beta_i \alpha_i} \right) - \frac{1 + \gamma}{2} \right]
\]
Since $\partial \beta_i^\# / \partial \alpha > 0$ and $\beta_i^\# = \alpha_i = 1/2$ at $\alpha_i = 1/2$ (by part (i) and symmetry), it is easy to check that $\partial L / \partial y > 0$ as $\alpha_i \to 0$, $\partial L / \partial y < 0$ as $\alpha_i \to 1$, and $\partial L / \partial y = 0$ for $\beta_i = \alpha_i = 1/2$. Thus by the continuity of $\partial L / \partial y$, we need to show that $\partial L / \partial y > 0$ if and only if $\beta_i = \alpha_i = 1/2$.

Writing $k = 2/1 + y$ and $x = ((1 - \beta_i)(1 - \alpha_i)/\beta_i \alpha_i)^k$, $\partial L / \partial y = 0$ can be expressed as follows:

$$\left( \frac{1 - \beta_i}{\beta_i} \right)^4 = \frac{x^2(1 - \ln x)}{1 + \ln x}$$

By Part (i), the LHS is a decreasing function of $\alpha_i$, while the RHS is an increasing function of $\alpha_i$, for

$$\frac{d}{dx} \left( \frac{x^2(1 - \ln x)}{1 + \ln x} \right) = -2x \left( 1 - \frac{1}{1 + \ln x} \right)^2 < 0 \quad \text{for} \quad x > 0$$

and $x$ is a decreasing function of $\alpha_i$ (by Part (i)).

Thus $\partial L / \partial y = 0$ has a single root $\alpha_i = 1/2$. It follows that $\partial L / \partial y > 0$ for $\alpha_i < 1/2$ and $\partial L / \partial y < 0$ for $\alpha_i > 1/2$ (Q.E.D.).

**Proof of Corollary 1.** It is easy to check that $\beta_i^\# = 0$ at $\alpha_i = 0$ and that $\beta_i^\# = 1$ at $\alpha_i = 1$. By Proposition 1, $\beta_i^\#$ is strictly increasing with $\alpha_i$, and thus for all $\alpha_i \neq 0, 1$, we have $0 < \beta_i^\# < 1$ (Q.E.D.).

Part (iii): The monitoring case.

**Proof of Proposition 2.** When person 1 monitors person 2, the optimal efforts are defined by the following three first-order conditions (corresponding to Eqs. (6) and (8) for $i = 2$ in the text).

\[
\frac{\partial f(a_1, \rho_1 b_1)}{\partial a_1} - (1 - \beta_1) \frac{\partial f(a_1, a_2)}{\partial a_1} - a_1 = 0
\]

\[
\rho_1 \frac{\partial f(a_1, \rho_1, b_1)}{\partial a_2} - b_1 - \rho_1^2 b_1 = 0
\]

\[
(1 - \beta_1) \frac{\partial f(a_1, a_2)}{\partial a_2} - a_2 = 0
\]

Differentiating the above three first-order conditions with respect to $\beta_1$ and $\rho_1$, respectively, and using Cramer’s rule, we have

\[
\frac{\partial a_1}{\partial \beta_1} = \frac{A}{D}, \quad \frac{\partial a_1}{\partial \rho_1} = \frac{B}{D}, \quad \text{and} \quad \frac{\partial b_1}{\partial \beta_1} = \frac{C}{D}
\]

\[25\] We owe James Mirrlees for his help in proving Part (ii).
where

\[ A = \left( \rho_1^2 \frac{\partial^2 f(a_1, \rho_1 b_1)}{\partial a_1^2} - (1 + \rho_1^2) \right) \]

\[ \times \left\{ \beta_2 - \frac{\partial^2 f(a_1, a_2)}{\partial a_1 \partial a_2} \left( \beta_2 \frac{\partial^2 f(a_1, a_2)}{\partial a_2^2} - 1 \right) - \frac{\partial f(a_1, a_2)}{\partial a_1} \right\} \]

\[ B = 2b_1 \frac{\partial^2 f(a_1, \rho_1 b_1)}{\partial a_1 \partial a_2} \left( \beta_2 \frac{\partial^2 f(a_1, a_2)}{\partial a_2^2} - 1 \right) \]

\[ C = \rho_1 \frac{\partial f(a_1, \rho_1 b_1)}{\partial a_1 a_2} \frac{\partial f(a_1, a_2)}{\partial a_2} \left( \beta_2 \frac{\partial^2 f(a_1, a_2)}{\partial a_2^2} - \beta_2 \frac{\partial^2 f(a_1, a_2)}{\partial a_1 \partial a_2} - 1 \right) \]

and

\[ D = \left( \beta_2 \frac{\partial^2 f(a_1, a_2)}{\partial a_2^2} - 1 \right) \left\{ \left( \frac{\partial f(a_1, \rho_1 b_1)}{\partial a_1^2} - \beta_2 \frac{\partial^2 f(a_1, a_2)}{\partial a_1^2} - 1 \right) \right. \]

\[ \times \left. \left( \frac{\partial f(a_1, \rho_1 b_1)}{\partial a_1^2} - (1 + \rho_1^2) \right) - \rho_1^2 \left( \frac{\partial f(a_1, \rho_1 b_1)}{\partial a_1 \partial a_2} \right)^2 \right\} \]

\[ + \beta_2^2 \left( \frac{\partial^2 f(a_1, a_2)}{\partial a_1 \partial a_2} \right)^2 \left( \frac{\partial f(a_1, \rho_1 b_1)}{\partial a_1^2} - (1 + \rho_1^2) \right) \]

It is easy to see that \( A < 0, B < 0 \) and \( C < 0 \). By using the properties of CES function, for \( \rho_1 b_1 > a_2 \), we have

\[ D \leq - \left( \rho_1^2 \frac{\partial^2 f(a_1, \rho_1 b_1)}{\partial a_2^2} - (1 + \rho_1^2) \right) \left( \beta_2 \frac{\partial^2 f(a_1, a_2)}{\partial a_2^2} - 1 \right) \]

\[ - \frac{\partial^2 f(a_1, a_2)}{\partial a_1^2} \left( \beta_2 \frac{\partial^2 f(a_1, \rho_1 b_1)}{\partial a_2^2} - (1 - \beta_2)(1 + \rho_1^2) \right) < 0 \]

Therefore \( \partial a_1 / \partial \beta_1 > 0, \partial a_1 / \partial \rho_1 > 0 \) and \( \partial b_1 / \partial \beta_1 > 0 \). By Eq. (9) in the text, total welfare also increases with \( \beta_1 \) and \( \rho_1 \) (Q.E.D.).

The proof of Corollary 2 is straightforward.

**Proof of Proposition 3.** When person 1 takes the full residual and monitors person 2, the optimal efforts are as follows:

\[ a_1^* = a_1 \left[ a_1 + (1 - a_1) \left( \frac{1 - \alpha_1 \rho_1^2}{\alpha_2 (1 + \rho_1^2)} \right)^{(1-\gamma)/(1+\gamma)} \right]^{\gamma/(1+\gamma)} \]
The total welfare is

\[ \Pi_1(\rho_1, 1) = \frac{1}{2} \alpha_1 \left[ \alpha_1 + (1 - \alpha_1) \left( \frac{1}{\alpha_1(1 + \rho_1^2)} \right)^{1/(1+\gamma)} \right]^{(1+\gamma)/(1-\gamma)} \]

Similarly, when person 2 takes the full residual and monitors person 1, the total welfare is

\[ \Pi_2(\rho_2, 1) = \frac{1}{2} \alpha_2 \left[ \alpha_2 + (1 - \alpha_2) \left( \frac{1}{\alpha_2(1 + \rho_2^2)} \right)^{1/(1+\gamma)} \right]^{(1+\gamma)/(1-\gamma)} \]

Now setting \( \Pi_2(\rho_1, 1) \geq \Pi_2(\rho_2, 1) \), we get inequality as Eq. (12) in the text (Q.E.D.). The proof of Corollary 3 is straightforward from Eq. (12).

**Proof of Proposition 4.** Part (i) has been proven in the text of Section 2 (by the intermediate value theorem). Parts (ii) and (iii) follow by the definition of \( \rho_i^* \) and Part (iii) of Proposition 2 (Q.E.D.).

**Appendix B. The Cobb–Douglas case**

The first-best solutions are

\[ a_i^{\text{FB}} = \alpha_i^{(1+\alpha_i)/2} / \alpha_j^{\alpha_j/2}, \quad i, j = 1, 2 \quad \text{and} \quad i \neq j \]

\[ y^{\text{FB}} = \alpha_1^{\alpha_1} \alpha_2^{\alpha_2} \]

\[ \Pi^{\text{FB}} = \frac{1}{2} \alpha_1^{\alpha_1} \alpha_2^{\alpha_2} \]

When there are no monitors, the partnership equilibrium is as follows

\[ a_i^# = (\alpha_i \beta_i)^{(1+\alpha_i)/2} (\alpha_j \beta_j)^{\alpha_j/2}, \quad i, j = 1, 2 \quad \text{and} \quad i \neq j \]

\[ y^# = (\alpha_1 \beta_1)^{\alpha_1} (\alpha_2 \beta_2)^{\alpha_2} \]

\[ \pi^# = \frac{1}{2} (\alpha_1 \beta_1)^{\alpha_1} (\alpha_2 \beta_2)^{\alpha_2} [1 + \beta_1 \alpha_2 + \beta_2 \alpha_1] \]

\[ \beta_i^# = \frac{\alpha_i (1 + \alpha_i) - \sqrt{\alpha_i (1 - \alpha_i^2 (2 - \alpha_i)}}}{2(2\alpha_i - 1)} \]

Clearly, the equilibrium efforts and output without monitoring are smaller than the first best efforts and output. In addition, output \( y^# \) is maximized when \( \beta_i = \alpha_i \).
If person 1 monitors person 2, the principalship equilibrium is as follows:

\[
\hat{a}_1 = a_1^{(1+\alpha_1)/2} a_2^{\alpha_2/2} \left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{\alpha_2/(1+\alpha_1)} - \beta_2^{2/(1+\alpha_1)} \left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{(1+\alpha_1)/2}
\]

\[
\hat{b}_1 = a_1^{\alpha_1/2} a_2^{(1+\alpha_2)/2} \left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{1/(1+\alpha_1)} - \beta_2^{2/(1+\alpha_1)} \left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{(1+\alpha_1)/2}
\]

\[
\hat{a}_2 = a_1^{\alpha_1/2} a_2^{(1+\alpha_2)/2} \rho_1^{1/(1+\alpha_1)} \left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{\alpha_2/(1+\alpha_1)} - \beta_2^{2/(1+\alpha_1)} \left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{(1+\alpha_1)/2}
\]

\[
\Pi_1(\rho_1, \beta_1) = \frac{1}{2} a_1^{\alpha_1} a_2^{\alpha_2} \left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{\alpha_2/(1+\alpha_1)} - \beta_2^{2/(1+\alpha_1)} \left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{(1+\alpha_1)/2}
\]

\[
\times \left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{\alpha_2/(1+\alpha_1)} + \alpha_1 \beta_2^{2/(1+\alpha_1)}
\]

Clearly, the equilibrium production effort and monitoring effort of person 1 and the output all increase with person 1’s share of output and effectiveness of monitoring. It is easy to see that \(\rho_1 \hat{b}_1 \geq \hat{a}_2\) if and only if \(\beta_2 \leq \rho_2^2/1 + \rho_2^2\). When \(\beta_2 = 0\), the inequality always holds. As \(\beta_2 = 0\) and \(\rho_1 \to \infty\), the principalship equilibrium converges to the first best allocation.

Similarly, when person 2 monitors person 1, we have

\[
\Pi_2(\rho_2, \beta_2) = \frac{1}{2} a_1^{\alpha_1} a_2^{\alpha_2} \left( \frac{\rho_2^2}{1 + \rho_2^2} \right)^{\alpha_1/(1+\alpha_2)} - \beta_1^{2/(1+\alpha_2)} \left( \frac{\rho_2^2}{1 + \rho_2^2} \right)^{(1+\alpha_2)/2}
\]

\[
\times \left( \frac{\rho_2^2}{1 + \rho_2^2} \right)^{\alpha_1/(1+\alpha_2)} + \alpha_2 \beta_1^{2/(1+\alpha_2)}
\]

It is easy to see \(\Pi_1(\rho_1, 1) \geq \Pi_2(\rho_2, 1)\) if and only if the following condition holds

\[
\left( \frac{\rho_1^2}{1 + \rho_1^2} \right)^{\alpha_2} \geq \left( \frac{\rho_2^2}{1 + \rho_2^2} \right)^{\alpha_1}
\]

As in Corollary 3, if \(\alpha_1 \geq \alpha_2\) and \(\rho_1 \geq \rho_2\), then \(\Pi_1(\rho_1, 1) \geq \Pi_2(\rho_2, 1)\).

References
