Essays on Microeconomics

By

Junjie Zhou

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Chris Shannon, Chair
Professor Robert Anderson
Professor Benjamin Hermalin
Professor Suzanne Scotchmer

Spring 2012

Essays on Microeconomics

Abstract

Essays on Microeconomics

by

Junjie Zhou

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor Chris Shannon, Chair

This dissertation consists of two essays on microeconomics.

The first chapter explores leadership within hierarchical organizations. For each hierarchy, I consider a dynamic signaling game in which each player observes only the actions of his direct superiors before choosing his action. At the top of the hierarchy are the leaders, who learn the state from nature. The hierarchy controls the flow of information and the timing of the game, and determines the equilibrium output and welfare. I show that the welfare-optimal hierarchy is the chain, because it maximizes the incentive of players to "lead by example" for their subordinates. The chain remains optimal even in the presence of verifiable or unverifiable costly information acquisition by the leaders. Lastly, I characterize optimal hierarchies when the number of layers or the number of leaders is limited. Applications to fund-raising are also discussed.

The second chapter studies the optimal way to select projects or agents in environments where information arrives in well defined rounds. Examples include academic environments where review periods are set by policy, aptitude tests such as those given by software developers to programmers applying for jobs, venture capital protocols where the rounds of funding may be stopped before the project is complete, and FDA testing, where drugs can be dropped at well defined junctures. Sequential rounds of elimination reduce the cost of selection, but also reduce the average quality of surviving projects. I characterize the nature of the optimal screening process with and without "memory." The second chapter is based on joint work with Suzanne Scotchmer.

To Fay

Contents

1 Economics of Leadership and Hierarchy **1**

1.1 Introduction 1

1.2 Model with Three Workers 5

1.2.1 Leading by Example 6

1.2.2 Sequential leading by example 7

1.2.3 Two leaders 10

1.2.4 Welfare comparison 12

1.3 The General Model 12

1.4 Simple Hierarchies 15

1.4.1 Welfare Comparisons for Simple Hierarchies 18

1.4.2 The Chain $\mathcal{C}_N$ 22

1.5 Complicated Hierarchies 24

1.5.1 Welfare Improving Operations 25

1.6 Extensions 27

1.6.1 Who will be a leader? 27

1.6.2 Limited Height 28

1.6.3 Endogenous Information Acquisition 30

1.6.4 Applications in fund-raising 33

1.7 Conclusion 34

2 Picking Winners in Rounds of Elimination **35**

2.1 Introduction 35

2.2 The model 37

2.3 Single Round of Elimination 39

2.3.1 Single Round: Promotion thresholds and monotonicity . . . 42

2.3.2 Examples 45

2.4 Double Rounds of elimination with memory 48

2.4.1 Double round with memory: monotonicity and threshold values 52

2.4.2 Double round with memory: comparisons 55

2.5 Double Rounds of elimination without memory 57

3 Appendix **63**

3.1 Some Auxiliary Lemmas 63

3.2 Omitted Proofs 68

3.3 Additional Materials 80

List of Tables

1.1 $\mathcal{C}_3$ with different shares 23

List of Figures

1.1 Standard Team Structure (T) 6
1.2 Leading by Example (Λ) 6
1.3 The Chain Structure (C_3) 8
1.4 V structure, with two leaders 10
1.5 Welfare Rankings 12
1.6 Examples of Hierarchies 14
1.7 Splitting 27
1.8 Individual payoffs with different hierarchies 28
1.9 two hierarchical structures in $\mathcal{M}^s(N, 3, 1)$ 29

2.1 selection on the mean 45
2.2 Selection for maximizing upside risk 47
2.3 Selection for minimizing downside risk 48

3.1 A counterexample: adding links is not always welfare improving. . . 80

Acknowledgments

There are many people to thank who helped me get through graduate school. First of all, I am indebted to my advisors Chris Shannon, Benjamin Hermalin and Suzanne Scotchmer for their nice advide, continuous guidance, great insight, and also patience. All of them are encouraging, gracious and helpful. I also thank: Robert Anderson, Ying-Ju Chen, Joan de Marti Beltran, Alessandro Pavan, Adam Szeidl and many speakers in the Economic Theory and Theory Lunch Seminars, for the beneficial discussions with them.

I am very grateful to Thomas Marschak for hundreds of discussions in the past four years and generous support in the summers of 2008-2012. I have learned a lot from you. Thank you, Tom.

Administrative supports from Barbara Waller, Patrick Allen, Marsha Snow and Barbara Peavy are highly appreciated. I would also like to thank my great friends: Shuchao Bi, Jianye He, An Huang, Jian Li, Yi Liu, Baoping Liu, John Zhu, and many others, for their encouragement and friendship.

I think my parents, Yuxuan Fan and Mingying Zhou, for their love and support. Last but not least, I would like to thank my wife Fang Huang for her love, her patience and her laughter.

Chapter 1

Economics of Leadership and Hierarchy

1.1 Introduction

This paper studies the role of leadership and information flow in the design of organizations. I develop a model of public good provision in teams with asymmetric information. Team members can engage in costly signaling of their information through their choice of effort to invest in the joint project. Leadership positions within the organization are distinguished by differential access to information: a team member's effort is observed only by her direct subordinates. The flow of information is thus endogenous to the design of the organization, and becomes the crucial channel through with the organizational design affects team output. I characterize the optimal organizational design in this model, and show that the optimal hierarchy provides important welfare gains over the standard team output and other methods of addressing the classic problem of moral hazard in teams (Holmstrom (1982)).

A central building block for my work is the idea of leading by example, introduced in the seminal work of Hermalin (1998). Hermalin (1998) also starts from the issue of free-riding in team production problems, and assumes that one team member knows the true marginal return to effort. In the standard team model, the informed member cannot credibly signal her information, thus it is useless. Hermalin's fundamental insight is that if the informed member can move first, however, then she can "lead by example": if she chooses her effort first and this is observable to all other team members, then her investment in the project provides a credible costly signal. Hermalin (1998) shows that such leading by example, by exploiting this information channel, yields higher welfare in equilibrium than the standard team production (even in the symmetric information case). Thus Hermalin (1998) identifies an important aspect of leadership in information transmission and incentive provision that can mitigate free-riding in teams. This insight has been at the heart of a sizable and growing literature on the economics of leadership (for example, see Hermalin (2007), Komai, Stegeman and Hermalin (2007), Komai and Stegeman (2010)).

An important limitation to the analysis in Hermalin (1998) is that the information channels and organizational structure are essentially taken to be exogenous. One team member is exogenously assumed to be informed about the true state, thus have "leadership potential." The leader's only choice is whether to move first, thereby signaling to all of the other members simultaneously. Thus the organizational structure is exogenously given: a two-tier hierarchy with the leader at the top and all other members on the second tier. If the signaling role of a leader is important, however, then information flows should be an important component in the endogenous design of organizations. For example, consider a three-person team in which one member learns the true state. Hermalin's (1998) results show that team output increases if the informed member invests first and reveals her investment to the other members, who then choose their investments simultaneously. Is this the optimal organizational design, however? Hermalin (1998) and the substantial work that followed do not address this important question. For example, is it better to have information flow through a "middle manager," that is, to have a three-tier hierarchy in which a single member observes the leader's investment, chooses his investment and then in turn reveals only his investment to the third member? Or is it better to have two leaders, each of whom signals to the third member?

To answer these questions, I start by illustrating the results in the simple case of three workers. In this case, it is possible to give an exhaustive list of all of the possible hierarchies. In a simplified version of the public good provision model with quadratic disutility of effort, I show that the optimal hierarchy has three tiers, with one leader on the top tier, one middle manager on the second tier, and a terminal worker on the third tier. This results in two rounds of the "leading by example" effect observed by Hermalin (1998). The middle manager "leads by example" for the terminal worker, which results in higher effort from the middle manager due to the need to provide credible signal. Because the middle manager works harder, the leader has a larger incentive to invest more as well, again due to the signaling effect. In particular, this three-tier chain hierarchy yields higher output than the two-tier hierarchy assumed in Hermalin (1998). Similarly, the chain dominates the inverted two-tier hierarchy with two leaders on the top tier.[1] In either case, one round of signaling is wasted, leading to lower output than in the chain.

[1] This result relies on an assumption on the beliefs of the terminal worker. See Section 2.3.

The general model with any number of workers and general sharing rule and disutility function is analyzed in sections 4 and 5. Here I distinguish between simple hierarchies, in which every player who is not the leader has a unique direct predecessor, and complicated hierarchies, in which at least one player has multiple direct predecessors. In a simple hierarchy, the dynamic signaling game I define always has a unique separating equilibrium, but in a complicated hierarchy typically there are multiple separating equilibria. Consequently, the analysis of complicated hierarchies is more delicate.

For simple hierarchies, I show that similar intuition as in the three-person example holds, and the chain is optimal in an arbitrarily large team. The optimality of the chain follows from the observation that by transforming any hierarchy into a chain, we obtain the maximal number of stages of signaling, as the set of followers for each member is larger in the chain than in the original hierarchy. For fixed shares, the chain gives every member the largest possible signaling incentive, hence motivates the highest efforts. Therefore, the chain can replicate the same welfare under any hierarchy but uses less total shares. Moreover, extra shares always improve welfare when distributed optimally among the team. Combining these results shows the optimality of the chain among simple hierarchies.

For complicated hierarchies, I focus on a particular equilibrium which shares a similar characterization as the unique equilibrium with simple hierarchies. Here I consider two operations on hierarchies: adding links and splitting. Adding links means constructing a link, and hence an information channel, between two members who were unconnected, while splitting means creating a new intermediate tier consisting of a single member chosen from a tier with more than one member, and adding the maximal number of links to this new tier. Interestingly, each operation improves welfare after adjusting the shares optimally. The optimality of the chain follows directly from the fact that any hierarchy can be transformed into a chain through a sequence of these two operations.

I then extend the model to allow for endogenous information acquisition by the leaders. If research effort is verifiable, then the optimal hierarchy is still the chain because the chain generates the highest social return to information. Even if research effort is not verifiable, the chain remains optimal because the leader's incentive for information acquisition now depends monotonically on her equilibrium effort, which is higher in the chain than in any other hierarchy. Thus the leader acquires more accurate information in equilibrium, even when research effort is not verifiable.

A drawback to the chain is that for a large team, the hierarchy is very long, as it requires as many tiers as team members. Thus I also consider a version of the model with constraints on the height of the hierarchy, that is, in which hierarchies are constrained to have fewer tiers than team members. In this case a chain is not feasible. I show that the optimal simple hierarchy must have the maximal number of middle managers, hence the smallest number of terminal workers. This is achieved by assigning at most one follower to each middle manager. The maximal number of middle managers exploits the maximal level of signaling incentives in the team when height is limited.

While I use the language of leaders and followers throughout the paper, following Hermalin (1998) and subsequent work, the results developed here can be applied to a wide variety of team production problems with asymmetric information. In many applications, the informed players who move first in the optimal team hierarchy need not literally be team "leaders" or CEOs; in many settings it might be natural for more informed members to be lower-level workers more familiar with the production technology or better able to collect information. In such problems, these results show that the optimal arrangement of the team is a chain originating with the informed member, with each member signaling via his effort to a subsequent member.

As an application, I consider the problem of a charity trying to raise funds from a pool of possible donors. I show that the charity can raise more money by implementing a fund-raising campaign resembling a chain; that is, by placing potential donors in a line and asking them to donate one after the other. In particular, the charity should not reveal the entire donation history to future donors.

Related Literature

This paper is related to two strands of literature, one focusing on the economics of leadership, and the other focusing on determinants of organizational design.

As mentioned an important contribution to the literature of leadership is Hermalin (1998), on which this paper is built. Many extensions of Hermalin's model have appeared. Komai and Stegeman (2010) study the rise of leaders endogenously. Komai, Stegeman, and Hermalin (2007) consider team production with binary action (participate or not). Hermalin (2008) extends the static model to an infinitely repeated setting, thus allowing the leader to build a reputation. The literature on leaders conveying information is surveyed in Hermalin (2007). As noted, a limitation of these models is that they all assume an exogenous organizational structure. The main contribution of this paper is to endogenize the organizational structure.

Many different aspects of hierarchies have been studied in previous work. Some approaches emphasize moral hazard and loss of control, for example, Calvo and Wellisz (1978, 1979), and Qian (1994). Others, following Radner (1993), study optimal hierarchies for minimizing costs of information processing and communication, for example, Bolton and Dewatripont (1994), Prat (1997), van Zandt

(1999), Marschak and Reichelstein (1998). This paper identifies another role of hierarchy through a different perspective, that of signaling channels. The dissemination of information along the hierarchy creates incentives for players as they try to influence their followers' beliefs, hence efforts.

This paper is also related to the increasingly growing literature on social learning (Bala and Goyal, 1998) and social networks (Jackson, 2008). The hierarchies studied here are particular networks in which information is transmitted through signaling along the directed links. In this model, cheap talk messages are not credible. Instead, the information about the state is transmitted from one player to his follower via his action, which connects this paper to the literature on "information cascades" and "herd behavior" (see, e.g., Banerjee 1992, Bikhchandani *et al.*, 1992). But there are many differences. First, followers in this paper have no private signals. Also, the action and state are both continua, the action fully reveals the state in equilibrium. Lastly, unlike that literature, one player's payoff here also depends on other players' actions.

1.2 Model with Three Workers

In this section, I provide basic intuition about how organizational structure affects the incentives of players and team welfare using a simplified three-worker public good production model. The general model with any fixed number of workers, general sharing rule and disutility function is studied in the next section.

For an organization with three workers, there are only a few possible hierarchies: T structure (team structure), Λ structure (leading by example), C_3 (sequential leading by example), and V structure (two leaders).[2] For each hierarchy, I define the game associated with it, characterize the unique separating equilibrium, and compute welfare in that equilibrium. I show the chain yields higher welfare than leading by example, which in turn yields higher welfare than the team structure. The analysis for the V structure is complicated by the fact that the unique follower of two leaders may have different out-of-equilibrium beliefs to support different equilibrium efforts from the leaders. Under a pessimistic belief assumption for the follower, I show that there is a continuum of separating equilibria, but all such equilibria are bounded by two special equilibria, what I call the U-equilibrium and L-equilibrium. The U-equilibrium does better than leading by example, but still worse than the chain, while the L-equilibrium does as well as the team structure. Hence, I give a complete picture of what we can achieve with three workers.

Consider a team with N=3 identical workers producing a joint project. The value of the project is $v(x_1, x_2, x_3) = \theta(x_1 + x_2 + x_3)$, where $\theta \in \Theta = [0, \infty)$

[2] There is another structure with two leaders, but one leader has no followers. See footnote 6 for discussion.

is a stochastic productivity factor and $x_i \in \mathbf{R}_+$ is the contribution of worker i. The prior distribution of θ is $F : [0, \infty) \to [0, 1]$. I assume F has full support and a continuous and positive density function f. Furthermore, assume each member gets 1/3 of the total output v and $c(x) = \frac{1}{2}x^2$ is the disutility of effort, which is the same for every worker. Then, worker i's payoff π_i is $\pi_i(x_1, x_2, x_3) = \frac{1}{3}v(x_1, x_2, x_3) - c(x_i) = \frac{1}{3}\theta(x_1 + x_2 + x_3) - \frac{1}{2}x_i^2$ and aggregate welfare is $W(x_1, x_2, x_3) = \sum_{i=1}^{3} \pi_i(x_1, x_2, x_3) = \sum_{i=1}^{3}(\theta x_i - \frac{1}{2}x_i^2)$. Note that the output function v is additively separable in individual efforts. So, if state θ is common knowledge or all workers have the same belief θ about the state, worker i has a dominant strategy to exert effort $x_i^N = \frac{1}{3}\theta$. Welfare under the corresponding equilibrium is $W^N(\theta) = W(\frac{1}{3}\theta, \frac{1}{3}\theta, \frac{1}{3}\theta) = \frac{5}{6}\theta^2$. The first-best effort is $x_i^{FB} = \theta$ and the first-best welfare is $W^{FB}(\theta) = W(\theta, \theta, \theta) = \frac{3}{2}\theta^2$. There is under-provision of effort due to standard free-riding in teams. In the first-best world, each worker must receive 100% of the output on the margin, but in total we only have 100% to give due to budget balance. The case with symmetric information can be graphically represented by Figure 1.1. Every member's position is symmetric in this structure, and we call it the standard team structure.

•1 •2 •3

Figure 1.1: Standard Team Structure (T)

1.2.1 Leading by Example

To counteract the free-riding problem in teams, hidden information and leading by example were introduced by Hermalin (1998). In that model, one worker, called the "leader," has superior information about θ and she moves first. All of the other workers, who initially only know the prior distribution of the state, observe her effort and choose their efforts simultaneously in the next stage.

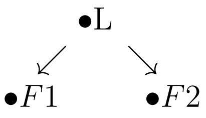


Figure 1.2: Leading by Example (Λ)

Figure 1.2 shows the relationship between workers. L is the leader and $F1$ and $F2$ are the two followers. An arrow from L to F1 means that F1 observes L's effort.

We call this the Λ structure since it resembles the letter Λ. For the Λ structure, we can define a signaling game as follows:

- Nature chooses $\theta \in \Theta$, which is unknown to all initially.
- In period 1, the leader L learns θ and chooses x_L.
- In period 2, both F1 and F2 observe x_L and pick their efforts x_{F1}, x_{F2} simultaneously.
- Payoffs are realized.

We are interested in separating equilibrium, in which the leader's effort $\tilde{x}_L(\theta)$ is a monotonic function of the state θ, hence reveals the state to the followers. A separating Perfect Bayesian Equilibrium (SPBE) of this signaling game is calculated in Hermalin (1998, Lemma 4). In that equilibrium, the leader's equilibrium strategy is $\tilde{x}_L(\theta) = \frac{2}{3}\theta$. After observing the leader's effort x_L, each follower's point belief about the state is $\beta_i(x_L) = \frac{3}{2}x_L, i = 1, 2$. Hence, each follower chooses effort $x_{F_i}(x_L) = \frac{1}{3}\beta_i(x_L) = \frac{1}{2}x_L, i = 1, 2$. The equilibrium efforts of $L, F1, F2$ are $\frac{2}{3}\theta$, $\frac{1}{3}\theta$, and $\frac{1}{3}\theta$ respectively.

The equilibrium welfare in this equilibrium is $W^L(\theta) = W(\frac{2}{3}\theta, \frac{1}{3}\theta, \frac{1}{3}\theta) = \theta^2$ So, the welfare with leading by example is higher than that with symmetric information, i.e, $W^L = \theta^2 > W^N = \frac{5}{6}\theta^2$.[3] Compared with the team structure, leading by example improves welfare, because the equilibrium effort of the leader is larger than under symmetric information, but still below the first-best level. The intuition is as follows. The leader gets a portion of the output generated by followers. In a separating equilibrium, the leader's effort fully reveals the information about the state. The harder the leader works, the higher the followers' beliefs about the state, thus the harder followers work and the better off is the leader. Being a leader and signaling to the followers gives leader L extra incentive to work hard beyond the incentive from her own share of the output. Given that the equilibrium efforts with symmetric information are too low to begin with, inducing harder work is welfare improving.

1.2.2 Sequential leading by example

Leading by example improves welfare in teams as shown by Hermalin (1998). But in equilibrium, efforts of the followers F1 and F2 are still too low. Welfare would be even higher if we could motivate any of the followers to exert higher effort in equilibrium. This requires that the followers have some extra signaling incentive as the leader L had in the previous example. This leads us to change the Λ structure

[3] This is true except when $\theta = 0$. For the sake of brevity, I will not repeat this caveat later.

into a chain (see Figure 1.3). The timing of the game is now modified as follows. In period 2, only F1 (not F2) can observe L's effort x_L, and he exerts effort in period 2. In period 3, F2 observes F1's effort x_{F1} and exerts effort x_{F2} last.

$$\bullet \mathrm{L} \rightarrow \bullet F1 \rightarrow \bullet F2$$

Figure 1.3: The Chain Structure (C_3)

The chain structure contains two stages of leading by example. Leader L signals to F1, and F1 signals to F2. Critically, F2 cannot observe the leader L's effort directly in period 2, otherwise F1 would have no signaling incentive because F2 would have already known the state from L's effort.

A separating Perfect Bayesian Equilibrium (SPBE) of this game is a strategy profile $\langle \tilde{x}_L(\cdot), \tilde{x}_{F1}(\cdot), \tilde{x}_{F2}(\cdot)\rangle$ and posterior point beliefs $\beta_{F1}(\cdot), \beta_{F1}(\cdot)$ such that:

$$(S) \quad \text{All } \tilde{x}_i, i \in \{L, F1, F2\}, \beta_j, j \in \{F1, F2\} \text{ are monotonic.}$$

$$(PL) \quad \forall\theta,\ \tilde{x}_L(\theta) \in \arg\max_{x_L \geq 0} \frac{\theta}{3}\left(x_L + \tilde{x}_{F1}(x_L) + \tilde{x}_{F2}(\tilde{x}_{F1}(x_L))\right) - \frac{1}{2}x_L^2.$$

$$(P1) \quad \forall x_L,\ \tilde{x}_{F1}(x_L) \in \arg\max_{x_{F1} \geq 0} \frac{\beta_{F1}(x_L)}{3}\left(x_L + x_{F1} + \tilde{x}_{F2}(x_{F1})\right) - \frac{1}{2}x_{F1}^2.$$

$$(P2) \quad \forall x_{F1},\ \tilde{x}_{F2}(x_{F1}) \in \arg\max_{x_{F2} \geq 0} \frac{\beta_{F2}(x_{F1})}{3}\left(x_L + x_{F1} + x_{F2}\right) - \frac{1}{2}x_{F2}^2.$$

and

$$(B) \quad \beta_{F1} = \tilde{x}_L^{-1}, \text{and } \beta_{F2} = \beta_{F1} \circ \tilde{x}_{F1}^{-1}$$

S says that each player's effort fully reveals his belief about the state under any history. B specifies the belief updating rule. PL, $P1$, and $P2$ are the usual perfection conditions: Each player is acting optimally given other players' beliefs and best responses.

It is easy to check that the following is a SPBE of this game:[4]

$$\begin{aligned} & & \tilde{x}_L(\theta) &= k_L\theta, \\ \beta_{F1}(x_L) &= \frac{x_L}{k_L} & \tilde{x}_{F1}(x_L) &= k_{F1}\beta_{F1}(x_L) = \frac{k_{F1}}{k_L}x_L \\ \beta_{F2}(x_{F1}) &= \frac{x_{F1}}{k_{F1}} & \tilde{x}_{F2}(x_{F1}) &= k_{F2}\beta_{F2}(x_{F1}) = \frac{k_{F2}}{k_{F1}}x_{F1} \end{aligned}$$

The constants $k_i, i = \{L, F1, F2\}$ are given by

$$k_{F2} = \frac{1}{3} \approx 0.333, \qquad k_{F1} = \frac{1+\sqrt{5}}{6} \approx 0.539, \qquad k_L = \frac{1+\sqrt{7+2\sqrt{5}}}{6} \approx 0.731$$

[4] This result is a special case of Theorem 4.1.

The equilibrium effort of player i is $k_i\theta, i = \{L, F1, F2\}$ and the corresponding welfare is

$$W^S \;=\; W(k_L\theta, k_{F1}\theta, k_{F2}\theta) \approx 1.13554\theta^2$$

It is clear that $W^S > W^N$, so we have established that the chain structure (Figure 1.3) with sequential leadership yields greater welfare than the Λ structure (Figure 1.2) with a single leader.

This result can be interpreted in the following way. In the chain structure, F1 plays leading by example with F2 as F2 infers the state from his effort. Due to signaling, F1's response as a function of his belief about the state must be steeper than what he chooses when he has no followers. Note that the leader L cannot influence the beliefs of player $F2$ directly as $F2$ cannot observe his effort, but indirectly L can influence the belief and effort of F2 through the intermediate player F1.

The leader L benefits from output generated by both F1 and F2. Her effort provides the information about the state. The greater her effort, the higher F1's belief, the greater F1's effort, the higher F2's belief, the higher F2's effort, and the better off is the leader. But the difference is now that F1's response given his belief β_1 is steeper under the chain than under the Λ structure. Therefore, to determine L's incentive for signaling, we can imagine that there is just one stage of signaling, as in Figure 1.2, with the share of F1 modified to $k_{F1} \approx 0.539 > 1/3$. This extra benefit increases the leader's equilibrium effort, thus we obtain $k_L > 2/3$. Since all the equilibrium efforts are still below first-best levels, we have obtained our welfare-comparison result.

At this time, it is worthwhile to check the role of additively separable production functions in this model. First, additive separability isolates the signaling effects from other forces possibly driven by strategic complements or substitutes between workers. If information is symmetric, the timing of moves is irrelevant as everyone has a dominant strategy, thus organizational structure does not play any role. When information is asymmetric, all the welfare comparison results must come from the signaling incentives. Second, additive separability greatly simplifies the computation of equilibrium. The strategic role of effort in this model is that it conveys valuable information about the state from one player to his followers. In the dynamic signaling game defined above, each worker's incentive consists of two parts. The first (direct) part is his share of the output. The second (indirect) part is strengthening of incentives of workers whose beliefs he can influence by his own effort, either directly or indirectly. In the chain, F1's effort can only affect F2's belief, while L can influence F1's belief directly and F2's belief indirectly. F2 cannot influence anyone's belief and therefore the second incentive component is zero. The higher is each incentive part, the harder one works.

1.2.3 Two leaders

Beyond the chain and the Λ structure, there is another hierarchy, called the V structure, possible with three workers. In the V structure, L1 and L2 are leaders and F is the only follower of both. See Figure 1.4. The time line for the V structure is the following:

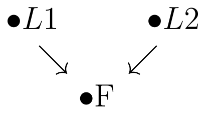


Figure 1.4: V structure, with two leaders

- Nature chooses $\theta \in [0, \infty)$, which is unknown to all.
- In period 1, both leaders L1 and L2 learn θ and choose x_1, x_2 simultaneously.
- In period 2, F observes both leaders' efforts x_1, x_2 and exerts effort x_F.
- Payoffs are realized.

A SPBE of this game is a strategy profile $\langle e_1(\cdot), e_2(\cdot), e_F(\cdot,\cdot)\rangle$ and belief function $b_F(\cdot,\cdot)$ such that:

$$\begin{array}{ll}(S) & e_i(\theta),\ i = \{1,2\} \text{ are monotonic.}\\ (P1) & \forall\theta,\ e_1(\theta) \in \arg\max\limits_{x_1\geq 0} \dfrac{\theta}{3}\left(x_1 + e_F(x_1, e_2(\theta)) + e_2(\theta)\right) - \dfrac{1}{2}x_1^2.\\ (P2) & \forall\theta,\ e_2(\theta) \in \arg\max\limits_{x_2\geq 0} \dfrac{\theta}{3}\left(e_1(\theta) + e_F(e_1(\theta), x_2) + x_2)\right) - \dfrac{1}{2}x_2^2.\\ (PF) & \forall x_1, x_2 \geq 0,\ e_F(x_1,x_2) \in \arg\max\limits_{x_f\geq 0} \dfrac{b_F(x_1,x_2)}{3}\left(x_1 + x_2 + x_f\right) - \dfrac{1}{2}x_f^2.\\ (B) & \forall\theta, b_F(e_1(\theta), e_2(\theta)) = \theta.\end{array}$$

P1 and P2 and PF are the perfection conditions. Note that B says that the belief of F is correct on the equilibrium path, but it is silent about F's out-of-equilibrium beliefs.

Obviously, PF could be replaced by

$$(PF') \qquad e_F(x_1, x_2) = b_F(x_1, x_2)/3$$

For convenience, define $\beta : \Theta \times \Theta \to \Theta$ by $\beta(\theta_1, \theta_2) := b_F\left(e_1(\theta_1), e_2(\theta_2)\right)$. Then $b_F(x_1, x_2) = \beta\left(e_1^{-1}(x_1), e_2^{-1}(x_2)\right)$. After simplifying and using (PF'), we can rewrite conditions P1 and P2 and B as:

$$
\begin{aligned}
&(P1') \quad \forall\theta,\ e_1(\theta) \in \arg\max_{x_1 \geq 0} \frac{\theta}{3}\left(x_1 + \frac{1}{3}\beta(e_1^{-1}(x_1), \theta)\right) - \frac{1}{2}x_1^2. \\
&(P2') \quad \forall\theta,\ e_2(\theta) \in \arg\max_{x_2 \geq 0} \frac{\theta}{3}\left(x_2 + \frac{1}{3}\beta(\theta, e_2^{-1}(x_2))\right) - \frac{1}{2}x_2^2. \\
&(B') \quad \forall\theta, \beta(\theta, \theta) = \theta.
\end{aligned}
$$

To find an equilibrium, it suffices to find functions $\{e_1, e_2, \beta\}$ satisfying conditions $S, P1', P2', B'$. There are many belief functions satisfying B'. I will use the following pessimistic belief assumption.

$$\beta^p(\theta_1, \theta_2) = \min(\theta_1, \theta_2). \tag{1.1}$$

There are two main justifications for this assumption. First, it gives both leaders some incentive to signal to the follower. If any of the leaders shirks, the follower will detect it immediately and punish that deviating leader by expending lower effort because the follower believes that the minimum of the two signals, here the one revealed by the deviating leader, is the true state. Second, if the follower's belief is the maximum of the two signals, then it is impossible to support any separating equilibrium, because at least one leader will have incentive to deviate either downward (free-riding on the other leader to signal to the follower) or upward (taking advantage of the benefits of signaling as the follower, in this case, will depend solely on his effort for updating beliefs).[5] If the follower believes that off the equilibrium path only one of the leaders is deviating, which is mostly likely to be the case, the pessimistic belief function is a better choice as the optimistic belief cannot support any equilibrium. Hence, we will maintain this pessimistic belief assumption from now on. For detailed discussions of the belief functions in the V structure, see the appendix.

Fix $\beta = \beta^p$. Let $\bar{e}(\theta) := \frac{1+\sqrt{5}}{6}\theta$ and $\underline{e}(\theta) := \frac{\theta}{3}$. It is easy to verify that $e_i = \bar{e}(\theta)$, $i = 1, 2$, or $e_i = \underline{e}(\theta)$, $i = 1, 2$ satisfy the above conditions $(P1'), (P2'), (B')$ with $\beta = \beta^p$. We call the equilibrium with $e_i = \bar{e}(\theta)$ $(\underline{e}(\theta))$, $\forall i$ the U(L)-equilibrium. In the U-equilibrium, both leaders exert higher efforts $\bar{e}$, while in the L-equilibrium, both leaders choose $\underline{e} = \frac{1}{3}\theta = x_i^N$, their efforts under symmetric information. In the appendix, I show that all separating equilibria corresponding to β^p are bounded by these two equilibria. Therefore, I can find the upper and lower bounds of the corresponding equilibrium welfare:

$$\frac{5}{6}\theta^2 \leq W^{2L} \leq \frac{(8 + 5\sqrt{5})}{18}\theta^2$$

[5] See appendix section C.2 for a detailed proof.

The lower bound corresponds to the L-equilibrium, while the upper bound corresponds to the U-equilibrium. Any number in between is also obtainable.

1.2.4 Welfare comparison

In previous subsections, we solve for equilibrium under four different structures with three workers. The computed welfares are ranked as follows:

$$W^N = \min W^{2L} < W^L < \max W^{2L} < W^S < W^{FB}$$

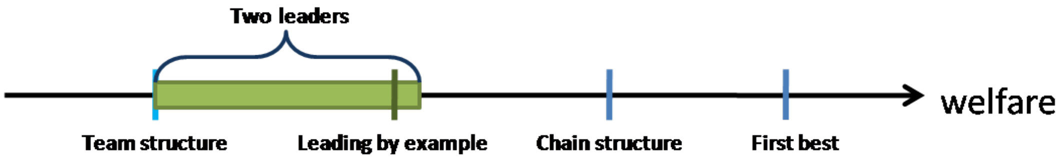


Figure 1.5: Welfare Rankings

As a conclusion, the chain structure dominates each of the other three structures.[6] None of them can achieve first-best welfare, however. Figure 1.5 summarizes these results.

As mentioned, workers on the higher tier of a chain exert higher efforts. One concern is that their efforts are inefficiently high, possibly in a team with large N. Part of the following analysis is to demonstrate that this phenomena is not going to happen in a team with arbitrary N. In the next section, I study the general model, and prove the optimality of the chain in the end.

1.3 The General Model

Let $\mathcal{N} = \{1, 2, \cdots, N\}$. Consider a team with N identical members. Each member n chooses an effort $e_n \in [0, \infty)$. The value to the team is $V = \theta \sum_{n=1}^{N} e_n$, where $\theta \in \Theta = [0, \infty)$ is a stochastic productivity factor. Each member has utility function $w - c(e)$, where w is his wage, and the disutility function c is twice differentiable and strictly convex with $c(0) = c'(0) = 0, c'(\infty) = \infty$.

Following Holmstrom (1982), we assume that contracts can only be written contingent on total output V, not on individual efforts. Furthermore, we restrict attention to affine-shares contracts, i.e., $w_n(V) = s_n V + t_n$. Here $\{s_n, t_n\}_{n=1}^{N}$

[6] Technically, there is one more structure with three workers constructed by deleting the link from L2 to F in the V structure. It is easy to see that this structure is less efficient than the Λ structure, hence worse than the chain.

are constants. No external source of funds means that $\sum_{n=1}^{N} w_n \leq V$. Requiring contracts to be renegotiation-proof means $\sum_{n=1}^{N} w_n \not< V$. Therefore $\sum_{n=1}^{N} s_n = 1, s_n \geq 0$, and $\sum_{n=1}^{N} t_n = 0$. We are interested in team members' equilibrium efforts. The transfers $\{t_n\}$ are irrelevant to our analysis, hence omitted in our calculations. Let $\Delta^N = \{\mathbf{s} = (s_1, \cdots, s_N) | \sum_{i=1}^{N} s_i = 1, s_i \geq 0, \forall i\}$ denote the N-dimensional simplex.[7] Then an affine-shares contract is just an element $\mathbf{s}$ of Δ^N.

As we have seen in section 2, hierarchy matters for the performance of the team. To define a hierarchy formally, we first review some concepts from graph theory.

Definition 1 (Graph) *A* **directed graph** $(\mathcal{N}, \mathbf{d})$ *consists of a set of nodes* $\mathcal{N} = \{1, 2, \cdots, N\}$ *and an adjacency matrix* $\mathbf{d} = (d_{ij})_{N\times N}$. $d_{ij} = 1$ *if there is a* **directed link** *from i to j, otherwise $d_{ij} = 0$.*
A **path** *from i to j is a sequence of nodes $i_1, i_2, \cdots, i_k$ such that $i_k = j$, and $d_{ii_1} = d_{i_2i_3} = \cdots d_{i_{k-1}i_k} = 1$, while k is called the* **length** *of this path.*

In this paper, $\mathcal{N}$ is fixed, thus we refer to $\mathbf{d}$ as a graph. For a fixed adjacency matrix $\mathbf{d}$, we can define $DF^i = \{j \in \mathcal{N} | d_{ij} = 1\}$ as the set of direct followers of i, and $F^i = \{j \in \mathcal{N} |\text{there is a path from } i \text{ to } j\}$ as the set of i's followers, direct and indirect. Similarly, we define $DP^i = \{j \in \mathcal{N} | d_{ji} = 1\}$ as the set of direct predecessors of i, and $P^i = \{j \in \mathcal{N} |\text{there is a path from } j \text{ to } i\}$ as the set of i's predecessors, direct and indirect. By definition, $j \in DF^i$ if and only if $i \in DP^j$. Similarly, $j \in F^i$ if and only if $i \in P^j$. Obviously $DF^i \subset F^i$ and $DP^i \subset P^i, \forall i$.

Definition 2 (Ordered Partition) $\mathcal{P} = \{A^1, A^2, \cdots, A^m\}$ *is called an* **ordered partition** *of A if (1) $A^i \neq \emptyset$ for all i; (2) $\cup_{i=1}^{m} A^i = A$; (3) $A^i \cap A^j = \emptyset, \forall i \neq j$.*

Definition 3 (Hierarchy) *A* **hierarchy** $\mathcal{H} = (\mathbf{d}, \mathcal{P})$ *on $\mathcal{N}$ is a directed graph $(\mathcal{N}, \mathbf{d})$ together with an ordered partition $\mathcal{P}= \{N^1, N^2, \cdots, N^h\}$ of $\mathcal{N}$ such that:*

(*a*) *for $k = 1, 2, \cdots, h-1$, if $j \in N^k$ and $d_{ji} = 1$, then $i \in N^{k+1}$. If $j \in N^h$, then $d_{ji} = 0, \forall i$.*

(*b*) *for $k = 2, \cdots, h$, if $j \in N^k$, then there is a node $i \in N^{k-1}$ such that $d_{ij} = 1$.*

Definition 4 *In a hierarchy $\mathcal{H} = (\mathbf{d}, \mathcal{P})$ with $\mathcal{P}= \{N^1, N^2, \cdots, N^h\}$, the number h is called the* **height** *of $\mathcal{H}$.*

[7] Later on, I will allow $\sum_{n=1}^{N} s_i \leq 1$ in some of the proofs.

For a fixed hierarchy $\mathcal{H}$, N^k is just the set of members on level k, and $n^k = |N^k|$ is the number of workers on level k. The workers in N^1 are called *the leaders*, while the workers in N^h are called *terminal workers.*

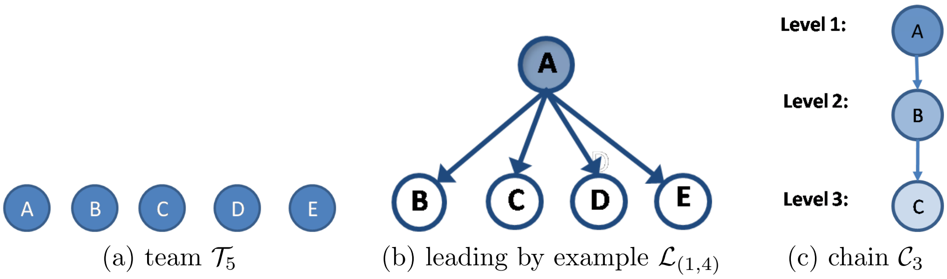
Condition (a) means that a path can only connect nodes from one level to the next lower level, while condition (b) says that all workers except leaders have at least one predecessor. Terminal workers have no followers and leaders have no predecessors.

Remark 1 *Notice that from Definition 3(a), direct links only follow from the leaders to middle level workers, and from middle level workers to terminal workers. The length of any path in a hierarchy is hence bounded above by the height of the hierarchy. Also, links between nodes on the same level are not allowed.*

When the partition is obvious from the context, I refer to the adjacency matrix $\mathbf{d}$ as the hierarchy. Here are some examples of hierarchies.

Example 1 ($\mathcal{T}_N$) *Figure 1.6a is a hierarchy with N=5, which is the standard team structure. The general team with N members is denoted* $\mathcal{T}_N$.

Example 2 ($\mathcal{L}_{(1,N-1)}$) *Figure 1.6b is a hierarchy with height 2, one level with one leader and the other level with 4 direct followers. This is the structure explored in Hermalin (1998). The general leading by example hierarchy with one leader and* $N-1$ *followers is denoted* $\mathcal{L}_{(1,N-1)}$.



(a) team $\mathcal{T}_5$ (b) leading by example $\mathcal{L}_{(1,4)}$ (c) chain $\mathcal{C}_3$

Figure 1.6: Examples of Hierarchies

Example 3 ($\mathcal{C}_N$) *Figure 1.6c is a hierarchy with three levels. Member B follows A, but is followed by C. The general chain with N workers is denoted* $\mathcal{C}_N$.

We have analyzed these examples in section 2 for the case $N = 3$. A common feature of the above three examples is that a worker has at most one direct predecessor, and hence can make inferences regarding the state only through that predecessor's effort. The equilibrium outcomes are quite different when a worker can draw multiple inferences of the state from efforts of his predecessors. I distinguish these two cases.

Definition 5 *A hierarchy $\mathcal{H}$ is called* **simple** *if any worker who is not a leader has a unique predecessor; $\mathcal{H}$ is called* **complicated** *if it is not simple.*

In Examples 1-3, hierarchies are all simple. The V structure (Figure 1.4) is a complicated hierarchy.

Given the payoff defined by the contract $\mathbf{s}$, and the timing defined by the hierarchy $\mathcal{H}$, I can study the equilibrium and make welfare comparisons across different hierarchies as I did in the previous section.

1.4 Simple Hierarchies

In this section we study the general case of simple hierarchies.

Given an affine-shares contract $\mathbf{s} = \{s_i, i \in \mathcal{N}\} \in \Delta^N$ and a simple hierarchy $\mathcal{H}$, define an $h+1$ stage dynamic game $G(\mathbf{s}, \mathcal{H})$ as follows:

- $t = 0$, nature chooses θ, which is unknown to all.
- $t = 1$, the leaders in N^1 learn θ and exert effort simultaneously.
- $t = 2$, each member $j \in N^2$ observes the effort of his unique direct predecessor DP^j(a singleton set in this case), who exerted effort in period $t = 1$. Then, members in N^2 exert effort simultaneously.

  $\vdots$

- $t = k + 1$, each member $l \in N^{k+1}$ observes the effort of his unique direct predecessor DP^l, who exerted effort in period $t = k$. Then, members in N^{1+k} exert effort simultaneously.

  $\vdots$

- θ is realized and output is divided according to $\mathbf{s}$.

The following theorem fully characterizes the separating equilibrium.

Theorem 1.4.1 (Equilibrium Characterization) *If the hierarchy $\mathcal{H}$ is simple, then there exists a separating equilibrium of $G(\mathbf{s}, \mathcal{H})$ in which the equilibrium efforts $\tilde{x}_i(\theta)$ are the solutions to the following system of differential equations:*

$$s_i\theta \left(1 + \sum_{j \ \in \ F^i} \frac{\tilde{x}'_j(\theta)}{\tilde{x}'_i(\theta)}\right) = c'(\tilde{x}_i(\theta)), \qquad i = 1, 2, \cdots, N. \tag{1.2}$$

If $F^i = \emptyset$, then the summation $\sum_{j \ \in \ F^i} \frac{\tilde{x}'_j(\theta)}{\tilde{x}'_i(\theta)}$ is zero by definition.

This result follows from a standard cost-benefit analysis. If player i deviates by exerting Δx_i more effort, then that effort will affect the beliefs of all his followers by $\Delta\theta \approx \frac{\Delta x_i}{\tilde{x}'_i(\theta)}$.[8] Hence each $j \in F^i$ will contribute more by the amount $\tilde{x}_j(\theta+\Delta\theta) - \tilde{x}_j(\theta) \approx \Delta\theta \cdot \tilde{x}'_j(\theta)$. Therefore the benefit of this deviation to i is approximately

$$s_i\theta \left(\Delta x_i + \sum_{j \in F^i} \Delta\theta \cdot x'_j(\theta)\right) \approx s_i\theta \left(1 + \sum_{j \ \in \ F^i} \frac{\tilde{x}'_j(\theta)}{\tilde{x}'_i(\theta)}\right) \Delta x_i$$

The cost of this deviation is:

$$c(\tilde{x}_i(\theta) + \Delta x_i) - c(\tilde{x}_i(\theta)) \approx c'(\tilde{x}_i(\theta))\Delta x_i$$

In equilibrium, the benefit equals the cost, which leads to Equation (1.2).

The equilibrium characterization from Theorem 4.1 has some interesting features. First, the equilibrium effort of player i only depends on his share s_i and the equilibrium efforts of his followers. It does not depend on the effort of his predecessors, or on the effort of other workers on the same level as him. Again, this follows from additive separability of the production function. Second, all followers, not just direct followers, affect equilibrium effort. Third, these equations are recursively solvable. Solving for a player's equilibrium effort function requires solving for all of his followers' equilibrium efforts, which requires solving for the equilibrium efforts of the followers' followers, etc.

The term $s_i\theta$ comes from i's share of the output, and the term $s_i\theta \sum_{j \ \in \ F^i} \frac{\tilde{x}'_j(\theta)}{\tilde{x}'_i(\theta)}$ comes from i's signaling incentive and it is always nonnegative. If the signaling term $\sum_{j \ \in \ F^i} \frac{\tilde{x}'_j(\theta)}{\tilde{x}'_i(\theta)}$ vanishes in equation 1.2, the solutions are just the efforts that workers would expend under symmetric information. With this extra signaling incentive, in equilibrium i must expend more effort. Formally, we have:

Proposition 1.4.2 *The equilibrium effort $\tilde{x}_i(\theta)$ of player i characterized in Theorem 4.1 is greater than his effort with symmetric information; that is,*

$$\tilde{x}_i\left(\theta\right) \geq c'^{-1}(s_i\theta) \qquad i = 1, 2, \cdots, N.$$

[8] If $\mathcal{H}$ is simple, a deviation by player i *only* affects his followers.

Strict inequality holds if $\theta > 0$ and $F^i \neq \emptyset$.

Proof The weak version is obvious. If $\theta > 0$ and $F^i \neq \emptyset$, then $(1+\sum_{j\ \in\ F^i} \frac{\tilde{x}'_j(\theta)}{\tilde{x}'_i(\theta)}) > 1$, hence $c'(\tilde{x}_i(\theta)) > s_i\theta$, or $\tilde{x}_i\,(\theta) > c'^{-1}(s_i\theta)$. ■

Finding closed-form solutions to the above system of equations (1.2), in general, is infeasible. To get explicit solutions, further restrictions on the disutility functions are needed.

Assumption C: The disutility function is $c(x) = \frac{1}{2}x^2$.

Theorem 1.4.3 *Under Assumption C, the solutions to equation (1.2) are $\tilde{x}_i(\theta) = k_i(\mathbf{s}, \mathcal{H})\theta$, where $\mathbf{k}(\mathbf{s}, \mathcal{H}) = \{k_i(\mathbf{s}, \mathcal{H}), i \in \mathcal{N}\}$ satisfies the following equations:*

$$k_i(\mathbf{s}, \mathcal{H}) = \frac{s_i + \sqrt{s_i^2 + 4s_i\left(\sum_{j\in F^i} k_j(\mathbf{s}, \mathcal{H})\right)}}{2}, \qquad i = 1, 2, \cdots N \tag{1.3}$$

Proof For brevity, I write $k_i(\mathbf{s}, \mathcal{H})$ as k_i. If $\tilde{x}_i(\theta) = k_i\theta$, then $\tilde{x}'_i(\theta) = k_i$, so equation (1.2) is equivalent to:

$$s_i\theta\left(1 + \sum_{j\ \in\ F^i} \frac{k_j}{k_i}\right) = c'(\tilde{x}_i(\theta)) = \tilde{x}_i(\theta) = k_i\theta, \quad i = 1, 2, \cdots, N \tag{1.4}$$

Canceling θ:

$$s_i\left(1 + \frac{\sum_{j\ \in\ F^i} k_j}{k_i}\right) = k_i, \quad i = 1, 2, \cdots, N$$

Solving this quadratic equation gives us $k_i = \frac{s_i+\sqrt{s_i^2+4s_i\left(\sum_{j\in F^i} k_j\right)}}{2}$. ■

Define $g(x, y) = \frac{x+\sqrt{x^2+4xy}}{2}, x > 0, y \geq 0$ to be the unique positive solution to

$$x\left(1 + \frac{y}{g(x, y)}\right) = g(x, y).$$

See Lemma 3.1.1 in the appendix for some useful properties of g. Then equation (1.3) can be rewritten as:

$$k_i(\mathbf{s}, \mathcal{H}) = g(s_i, \sum_{j\in F^i} k_j(\mathbf{s}, \mathcal{H})), \quad i = 1, 2, \cdots, N \tag{1.5}$$

Theorem 1.4.3 gives linear solutions for equation (1.2) under quadratic disutility. The constant k_i, which quantifies how player i responds to his belief about the state in equilibrium, is called i's *responsive coefficient.* These coefficients can be calculated by equation 1.3 recursively.[9]

From Theorem 4.1 and Theorem 4.3, we see that each player's equilibrium effort depends positively on two components: one is the worker's share s_i of the total output, the other is the signaling part depending on the responsive coefficients of i's followers. For two workers with the same shares, if one worker is the follower of the other, then the signaling incentive of that follower should be weaker than the leader. It is intuitive to guess that the equilibrium effort of the leader should be higher than the follower. This result is formally presented in the following proposition.

Proposition 1.4.4 *If $s_i = s_j > 0$ and $j \in F^i$, then $\tilde{x}_i(\theta) > \tilde{x}_j(\theta)$.*

Proof If $j \in F^i$, then $F^j \subset F^i$. Therefore $k_i(\mathbf{s}, \mathcal{H}) > k_j(\mathbf{s}, \mathcal{H})$ by Theorem 1.4.3, hence $\tilde{x}_i(\theta) > \tilde{x}_j(\theta)$. ■

As a special case, for equal shares, we have:

Corollary 1.4.5 *If $s_i = 1/N, \forall i$, then the higher a player is in the hierarchy, the larger the equilibrium effort, and the smaller his equilibrium payoff.*

Proof If $s_i = s_j = 1/N$ and $j \in F^i$, then $\tilde{x}_i(\theta) > \tilde{x}_j(\theta)$ by Proposition 1.4.4, therefore $c(\tilde{x}_i(\theta)) > c(\tilde{x}_j(\theta))$. Then i works harder than j, but gets the same share of the output as j, so i's equilibrium payoff must be smaller. ■

1.4.1 Welfare Comparisons for Simple Hierarchies

In the previous subsection, I solved for the equilibrium efforts for a fixed hierarchy. The aggregate welfare definitely depends on the members' shares, but it also depends crucially on the structure of the hierarchy: how many members are on each level, how are they connected with each other, how many direct or indirect followers each worker has, and what the followers' shares are.

[9] Here is a simple algorithm to compute all the $\{k_i\}_{i=1}^N$ in N steps:

1. Start with terminal workers $j \in N^h$. Notice that $k_j = s_j$ for these workers.
2. Suppose we have computed k_j for all workers in $N^h, N^{h-1}, \cdots, N^k$. Then we can calculate k_i for each $i \in N^{k-1}$ using $k_i = g(s_i, \sum_{j \in F^i} k_j)$, since $F^i \subset \cup_{k \le t \le h} N^t$.

The aggregate welfare in the equilibrium characterized in Theorem 4.3 is

$$\begin{aligned} SW(\mathbf{s},\mathcal{H}) &= \theta\sum_{i\in\mathcal{N}} k_i(\mathbf{s},\mathcal{H})\theta - \sum_{i\in\mathcal{N}}\frac{1}{2}(k_i(\mathbf{s},\mathcal{H})\theta)^2 \\ &= \theta^2\sum_{i\in\mathcal{N}}\left(k_i(\mathbf{s},\mathcal{H}) - \frac{k_i(\mathbf{s},\mathcal{H})^2}{2}\right) \qquad (1.6) \end{aligned}$$

Clearly, the value of θ is irrelevant if we want to maximize aggregate welfare with respect to the hierarchy structure and the shares. Equivalently we define $w(\mathbf{s},\mathcal{H}){:=}\ \sum_{i\in\mathcal{N}}\left(k_i(\mathbf{s},\mathcal{H}) - \frac{k_i(\mathbf{s},\mathcal{H})^2}{2}\right)$. Then the welfare maximization problem can be written as

$$\max_{\mathbf{s},\mathcal{H}} w(\mathbf{s},\mathcal{H}) \quad \text{subject to } \sum_{i\in\mathcal{N}} s_i = 1, s_i \geq 0. \qquad (1.7)$$

This program can be decomposed into two steps:

1. For a fixed hierarchy $\mathcal{H}$, find the optimal shares $\mathbf{s}^*(\mathcal{H})$ by solving the following problem:

$$\max_{\mathbf{s}\geq\mathbf{0},\sum_{i\in\mathcal{N}} s_i=1} w(\mathbf{s},\mathcal{H}) = \sum_{i\in\mathcal{N}}\left(k_i - \frac{k_i^2}{2}\right) \qquad (1.8)$$

   We define $\bar{w}(\mathcal{H}) := w(\mathbf{s}^*(\mathcal{H}),\mathcal{H})$ as the maximum value above.

2. Maximize over different simple hierarchies with optimal shares $\mathbf{s}^*(\mathcal{H})$:

$$\max_{\mathcal{H}} w(\mathbf{s}^*(\mathcal{H}),\mathcal{H}) = \max_{\mathcal{H}} \bar{w}(\mathcal{H})$$

Below is the main result of this paper.

Theorem 1.4.6 (Optimal Simple Hierarchy) *The chain is the optimal simple hierarchy, i.e.,*

$$\bar{w}(\mathcal{H}) = \max_{\mathbf{s}\geq\mathbf{0},\sum_{i\in\mathcal{N}} s_i=1} w(\mathbf{s},\mathcal{H}) \leq \bar{w}(\mathcal{C}_N) = \max_{\mathbf{s}\geq\mathbf{0},\sum_{i\in\mathcal{N}} s_i=1} w(\mathbf{s},\mathcal{C}_N)$$

The proof is shown in a sequence of steps.

First, to make two hierarchies comparable, we transform any hierarchy $\mathcal{H}$ into a chain, and compare the equilibrium efforts under these two hierarchies using the same share profile. Each transformation is determined by a permutation σ of $\mathcal{N}$ by assigning member $\sigma(i)$ to the ith level on a chain, $i = 1,\cdots,N$. For a fixed hierarchy $\mathcal{H}$, we look for a special permutation σ satisfying the following condition:

Condition OP: If j is i's follower under $\mathcal{H}$, then $\sigma(i) < \sigma(j)$.

The existence of such a permutation is shown in the following lemma.[10]

[10] There are multiple permutations satisfying Condition OP. We just need one.

Lemma 1.4.7 *For any hierarchy $\mathcal{H}$, there exists a permutation σ satisfying Condition OP.*

Proof For each i, let $f_i = \#F^i$ be the number of followers of i under $\mathcal{H}$. We can enumerate the numbers in $\mathcal{N}$ by decreasing order of f_i (take any order if $f_i = f_j$ for $i \neq j$). We define the permutation σ by mapping each i to its place under this enumeration. Then, if $j \in F^i$, j's follower is also i's follower, hence F^j is a strict subset of F^i. Therefore $f_i > f_j$, so i must come earlier in σ than j, equivalently, $\sigma(i) < \sigma(j)$. ■

With the permutation σ given in the above lemma, we construct the chain $\mathcal{C}^{\sigma}_{(\mathcal{N})}$ by assigning member $\sigma(i)$ to the ith level, $i = 1, \cdots, N$. We call this the **C-transformation**, because this procedure transforms any hierarchy into a chain. Given a share profile $\mathbf{s}$ for $\mathcal{H}$, we assign the same share s_i to member i in the chain $\mathcal{C}^{\sigma}_{(\mathcal{N})}$. Condition OP implies that this transformation has some nice properties.

Lemma 1.4.8 *The C-transformation given by a permutation σ satisfying condition OP has the following properties.*

1. *For fixed shares $\mathbf{s}$, any member's equilibrium effort is weakly higher in $\mathcal{C}^{\sigma}_{(\mathcal{N})}$ than in $\mathcal{H}$, that is,*
$$k_i(\mathbf{s}, \mathcal{C}^{\sigma}_{(\mathcal{N})}) \geq k_i(\mathbf{s}, \mathcal{H})$$
2. *For fixed shares $\mathbf{s}$, we can find another share profile $\tilde{\mathbf{s}}$ such that $\tilde{\mathbf{s}} \leq \mathbf{s}$ but*
$$k_i(\tilde{\mathbf{s}}, \mathcal{C}^{\sigma}_{(\mathcal{N})}) = k_i(\mathbf{s}, \mathcal{H})$$
Therefore, $w(\tilde{\mathbf{s}}, \mathcal{C}^{\sigma}_{(\mathcal{N})}) = w(\mathbf{s}, \mathcal{H})$.

The C-transformation preserves the subordination relation between members. If i is a predecessor of j in $\mathcal{H}$, then i is j's predecessor in $\mathcal{C}^{\sigma}_{(\mathcal{N})}$. For some players, however, the set of followers may be strictly larger in $\mathcal{C}^{\sigma}_{(\mathcal{N})}$. Each worker's equilibrium effort depends on his own share and the sum of equilibrium efforts of all his followers. Since the shares are the same under the two hierarchies, each member will have larger incentive to signal when he has more followers in $\mathcal{C}^{\sigma}_{(\mathcal{N})}$, therefore his equilibrium effort is weakly higher. Moreover we can reduce his share suitably to make his equilibrium effort equal under two hierarchies. Therefore for any hierarchy $\mathcal{H}$ with a share profile $\mathbf{s}$, the chain $\mathcal{C}^{\sigma}_{(\mathcal{N})}$ can generate the same welfare using a share profile $\tilde{\mathbf{s}}$, which uses less total shares than $\mathbf{s}$. But $\tilde{\mathbf{s}}$ in general does not belong to Δ^N even if the initial share profile $\mathbf{s}$ is in Δ^N.

Next we show that if we distribute the the extra shares optimally to the team members in the chain, we can generate even higher welfare. Actually we show this result for any hierarchy, not just the chain.

Formally, for each fixed $\mathcal{H}$ and $t \in (0,1]$, define

$$\begin{aligned}\phi(t,\mathcal{H}) := \quad & \max_{\mathbf{s}} w(\mathbf{s},\mathcal{H}) \\ & \text{s.t:} \sum_{i\in\mathcal{N}} s_i = t, s_i \geq 0\end{aligned}$$

as the maximal achievable welfare under $\mathcal{H}$ with the constraint that the total shares sum up to t. In particular when $t = 1$, $\phi(1,\mathcal{H}) = \bar{w}(\mathcal{H})$. In the appendix, we show that increasing the total constraint on shares improves welfare whenever total shares are less than 1.

Theorem 1.4.9 *For any simple hierarchy $\mathcal{H}$, if $1 \geq t_1 > t_2 \geq 0$, then $\phi(t_1,\mathcal{H}) > \phi(t_2,\mathcal{H})$.*

Using Lemma 1.4.8 and Theorem 1.4.9, we show that the chain structure is optimal whenever the total shares are less than 1.

Proposition 1.4.10 *For any simple hierachy $\mathcal{H}$, $\phi(t,\mathcal{H}) \leq \phi(t,\mathcal{C}_N), \forall t \in (0,1]$.*

Proof For any $t \in (0,1]$, suppose $\mathbf{s}$ is optimal, i.e., $\phi(t,\mathcal{H}) = w(\mathbf{s},\mathcal{H})$. Then from Lemma 1.4.8, we can find a welfare equivalent $\tilde{\mathbf{s}}$ in the chain $C^{\sigma}_{(N)}$, but use less total share, i.e, $w(\mathbf{s},\mathcal{H}) = w(\tilde{s},\mathcal{C}^{\sigma}_{(\mathcal{N})})$, $|\tilde{\mathbf{s}}| \leq |\mathbf{s}| = t$. By Theorem 1.4.9, extra shares always yield greater welfare if we adjust the shares optimally. Formally, we have:

$$\begin{aligned}\phi(t,\mathcal{H}) &= w(\mathbf{s},\mathcal{H}) = w(\tilde{s},\mathcal{C}^{\sigma}_{(\mathcal{N})}) \\ &\leq \max_{\mathbf{s}\geq 0, |s| = |\tilde{\mathbf{s}}|} w(s,\mathcal{C}^{\sigma}_{(\mathcal{N})}) = \phi(|\tilde{\mathbf{s}}|,\mathcal{C}^{\sigma}_{(\mathcal{N})}) \leq \phi(t,\mathcal{C}^{\sigma}_{(\mathcal{N})})\end{aligned}$$

The first inequality holds since $\tilde{\mathbf{s}}$ is not necessarily optimal. The second inequality follows from Theorem 1.4.9 since $|\tilde{\mathbf{s}}| \leq |\mathbf{s}| = t$. Note that both $\mathcal{C}^{\sigma}_{(\mathcal{N})}$ and $\mathcal{C}_N$ represent the same hierarchy: the chain of length N. Therefore, $\phi(t,\mathcal{C}^{\sigma}_{(\mathcal{N})}) = \phi(t,\mathcal{C}_N)$.[11] Thus, we finish the proof of Proposition 1.4.10. ■

Proposition 1.4.10 immediately implies Theorem 1.4.6 by setting $t = 1$.

The optimality of the chain comes from three observations. First, the chain gives every member the maximal signaling incentives. Note that motivating efforts alone is not always welfare improving, as the welfare function is declining after effort exceeds first-best level. Second, we can use fewer shares to provide incentives to the workers in the chain to generate the same levels of efforts than in any other hierarchy. Third, extra shares, if distributed optimally among the team, improve welfare.

[11] This result does not hold if the team is not homogeneous either because different workers have different disutility functions, or because different workers value the project differently.

Remark 2 *If the main program is to maximize equilibrium output or equivalently $\sum_{i\in\mathcal{N}} k_i(\mathbf{s},\mathcal{H})$ rather than equilibrium welfare, then the chain is still optimal, since each worker's equilibrium effort is weakly higher after C-transformation (see Lemma 1.4.8).*

1.4.2 The Chain $\mathcal{C}_N$

Given the optimality of the chain, we next study the optimal sharing rule and equilibrium effort levels in the chain. We give a detailed analysis of the chain in this subsection.

Consider the chain $\mathcal{C}_N$. We denote by i the unique worker on level i, $i = 1,\cdots,N$. Let s_i be worker i' share. Then by Theorem 4.3, we have the following expression for $k_i(\mathbf{s},\mathcal{C}_N)$:

$$\begin{aligned}
k_N(\mathbf{s},\mathcal{C}_N) &= g\left(s_N,0\right) = s_N \\
k_{N-1}(\mathbf{s},\mathcal{C}_N) &= g\left(s_{N-1},k_N(\mathbf{s},\mathcal{C}_N)\right) \\
&\vdots \\
k_1(\mathbf{s},\mathcal{C}_N) &= g\left(s_1,\sum_{j=2}^{N} k_j(\mathbf{s},\mathcal{C}_N)\right)
\end{aligned}$$

Equal shares

To get quantitative results about the strength of sequential signaling effects—and also their limitations—here we calculate the welfare function for equal share $\mathbf{s}^{eq}$, that is, $s_i = \frac{1}{N},\forall i$. Equal shares corresponds to the case of public good provision or committee service, as each team member has roughly the same stake in the project.

First, we have the following estimates regarding responsive coefficients:

Lemma 1.4.11 *For equal shares $\mathbf{s}^{eq}$ in the chain $\mathcal{C}_N$, we have*

$$\frac{1}{2N} < k_i(\mathbf{s}^{eq},\mathcal{C}_N) - k_{i+1}(\mathbf{s}^{eq},\mathcal{C}_N) < \frac{1}{N} \qquad i = 1,\cdots,N-1$$

Hence:

$$\frac{N+1-i}{2N} < k_i(\mathbf{s}^{eq},\mathcal{C}_N) < \frac{N+1-i}{N} \qquad i = 1,\cdots,N-1$$

In particular, $k_i(\mathbf{s}^{eq},\mathcal{C}_N) < 1,\forall i$.

With these bounds, we can estimate the welfare $SW(\mathbf{s}^{eq},\mathcal{C}_N)$.

Proposition 1.4.12 *The aggregate welfare for $\mathcal{C}_N$ with equal shares, $SW(\mathbf{s}^{eq}, \mathcal{C}_N)$, satisfies:*

$$\frac{5}{24}N\theta^2 \leq SW(\mathbf{s}^{eq}, \mathcal{C}_N) \leq (\frac{1}{3}N + \frac{1}{4})\theta^2$$

The first-best aggregate welfare is $\frac{1}{2}N\theta^2$, which is, of course, infeasible here. Also, the free-riding problem is more severe in larger teams as each member's share is smaller. However, Proposition 1.4.12 shows that the signaling incentives in the team are strong enough to yield at least $\frac{5}{12} \approx 42\%$ of the first-best welfare with equal shares. Optimal shares potentially could do better.

From Lemma 1.4.11, we see that for the equal share rule, all the k_i are less the first-best. From Example 4 (below), however, we see that k_i could be greater than the first-best under some shares. So, we look for a sufficient condition on $\mathbf{s}$ to guarantee that $k_i(\mathbf{s}, \mathcal{C}_N) \leq 1$. The following proposition is one result along this line.

Proposition 1.4.13 *If the shares* $\mathbf{s}$ *for the chain are monotonic, so that* $0 < s_1 \leq s_2 \cdots \leq s_N$*, then* $k_i(\mathbf{s}, \mathcal{C}_N) < 1, \forall i$.

The assumption of this proposition is quite general. Equal shares satisfy this condition. By continuity, this conclusion still holds for shares s sufficiently close to s^{eq} or any monotonic shares. Note that $\mathbf{s}' = (0.8, 0.1, 0.1)$ in Example 4 is not monotonic, and indeed $k_A = 1.007$ is greater than 1 for $\mathbf{s}'$.

Optimal shares

In this subsection, we characterize the optimal shares for the chain with N workers. First, we start with an example with three workers.

Example 4 *Table 1.1 lists the equilibrium efforts and welfare with different share rules, where A is the leader, C is the terminal worker, and B is the middle worker.* $\mathbf{s}^{eq}$ *are the equal shares, and* $\mathbf{s}^*$ *are the optimal shares.*

$\mathbf{s} = (s_A, s_B, s_C)$	$\mathbf{k} = (k_A, k_B, k_C)$	welfare $w(\mathbf{s}, \mathcal{C}_3)$
$\mathbf{s}'$ =(0.8, 0.1, 0.1)	(**1.0078**, 0.1618, 0.1000)	0.7437
$\mathbf{s}''$ =(0.75, 0.1, 015)	(0.9994, 0.1823 ,0.1500)	0.8044
$\mathbf{s}^{eq}$ = $(0.3333, 0.3333, 0.3333)$	(0.7312, 0.5393, 0.3333)	1.1355
$\mathbf{s}^*$ =(0.1997, 0.2668, 0.5335)	(0.5721, 0.5335, 0.5335)	1.1909

Table 1.1: $\mathcal{C}_3$ with different shares

There are a few points worth noting. Because $k_A(\mathbf{s}', \mathcal{C}_3) > 1$, A's equilibrium effort is actually higher than first-best level. Transferring a small share from A to C, as shown in $\mathbf{s}''$, is welfare improving, because it mitigates A's overly strong signaling incentive in $\mathbf{s}'$ and gives C more shares so that C will expend higher effort. For equal shares $\mathbf{s}^{eq}$, efforts are monotonic, but still below first-best levels (Lemma 1.4.11). But equal shares are not optimal, i.e., $\mathbf{s}^{eq} \neq \mathbf{s}^*$. Moreover, for optimal shares, $s_A^* < s_B^* < s_C^*$, but $k_A^* > k_B^* = k_C^*$. The worker higher in the hierarchy actually has a smaller share, but works harder due to stronger signaling incentives from his followers. We will see that this is the general pattern for the chain of any length.

Equal shares are in general not optimal. The effort of workers on higher levels is too high compared with workers on lower levels. Now we turn to the question of how to find the optimal shares. Unfortunately, the expressions for k_i are recursive and the exact expressions involving $\{s_i\}$ are quite complicated. Therefore the Lagrange multiplier approach to maximize $w(\mathbf{s}, \mathcal{C}_N)$ with constraint $\mathbf{s} \in \Delta^N$ is not quite informative. Nevertheless, we use a variation argument to show that optimal shares $\mathbf{s}^*$ and the corresponding responsive coefficients $k_i^* = k_i(\mathbf{s}^*, \mathcal{C}_N)$ satisfy the following conditions:

Theorem 1.4.14 *The optimal shares* $\mathbf{s}^*$ *for the chain satisfy*

$$\frac{1}{2} < \frac{s_i^*}{s_{i+1}^*} < 1, \quad \forall i = 1, 2, \cdots, N-2, \quad \textit{and} \quad \frac{s_{N-1}^*}{s_N^*} = \frac{1}{2}.$$

Moreover, the k_i^**s satisfy* $1 > k_1^* > k_2^* > \cdots > k_{N-1}^* = k_N^* > 0$.

As a consequence, we get a chain of inequalities:

$$0 < s_1^* < s_2^* < \cdots < s_N^* = k_{N-1}^* = k_N^* < \cdots < k_2^* < k_1^* < 1. \tag{1.9}$$

All workers are symmetric ex ante, but we do not want to distribute the shares equally to them as different workers have different signaling incentives that vary with the their positions on the chain. Under optimal shares in the chain, a worker has stronger signaling incentives than his followers, although he gets a smaller share of the output.

1.5 Complicated Hierarchies

Although simple hierarchies are the structures typically observed in organizations, complicated hierarchies exist as well. Sometimes a worker may communicate with multiple bosses. I explore complicated hierarchies in this section.

For a complicated hierarchy, we can define the dynamic signaling game as before, except that now a worker can potentially observe efforts from multiple direct predecessors. After adopting the pessimistic belief assumption as I did for the V structure with two leaders, the equilibrium characterization is quite similar to the case of simple hierarchies.

Theorem 1.5.1 *If $\mathcal{H}$ is complicated, then there exists a separating equilibrium of $G(\mathbf{s}, \mathcal{H})$ in which the equilibrium efforts are $\tilde{x}_i(\theta) = k_i(\mathbf{s}, \mathcal{H})\theta$, while $k_i(\mathbf{s}, \mathcal{H})$ is given recursively by*

$$k_i(\mathbf{s}, \mathcal{H}) = g(s_i, \sum_{j \in F^i} k_j(\mathbf{s}, \mathcal{H})), \quad i = 1, 2, \cdots N.$$

Qualitatively, Theorem 1.5.1 looks exactly the same as Theorem 4.3. There are some major differences. First, we need the pessimistic belief assumption in complicated hierarchies, while we did not need any assumption on the belief functions for simple hierarchies. Second, there are multiple equilibria even with the pessimistic belief assumption as we have seen in the V structure (Section 2). Uniqueness of the equilibrium is not guaranteed.

From now on, we focus on the equilibrium identified by Theorem 1.5.1 for both simple and complicated hierarchies. Given the similarity between Theorem 1.5.1 and Theorem 4.3, we can show that the results proved for simple hierarchies, like Proposition 1.4.4, Lemma 1.4.8, and Theorem 1.4.9, also hold for complicated hierarchies. I will directly use them without proof.

1.5.1 Welfare Improving Operations

In section 4.1, we have shown that the chain structure is optimal among simple hierarchies. Next we want to show that the chain structure is optimal among all hierarchies, simple or complicated. To achieve this goal, we explore two welfare-improving operations on hierarchies.

Definition 6 *For fixed hierarchy $\mathcal{H} = (\mathbf{d}, \mathcal{P})$, let $\mathcal{H} + ij$ be the hierarchy formed from $\mathcal{H}$ by adding one direct link from i to j in $\mathbf{d}$.*

Note that in the definition i and j should not be linked in $\mathcal{H}$, and j should lie on the next level from i, otherwise $\mathcal{H} + ij$ is not a hierarchy according to Definition 3. Obviously, $\mathcal{H} + ij$ is not simple even if $\mathcal{H}$ is. Adding links enlarges the sets of followers, hence pushes workers' effort higher by Theorem 1.5.1. Formally:

Proposition 1.5.2 *For fixed share profile $\mathbf{s}$, $k_l(\mathbf{s}, \mathcal{H}) \leq k_l(\mathbf{s}, \mathcal{H} + ij)$, $\forall l \in N$.*

From Proposition 1.5.2, it is intuitive to guess that for fixed share profile $\mathbf{s}$, adding links will increase welfare: $w(\mathbf{s}, \mathcal{H}) \leq w(\mathbf{s}, \mathcal{H} + ij)$. We give a counterexample in the Appendix to show that this naive argument is wrong. The reason is that some players' effort is already too high under strong signaling incentives, and motivating higher effort for those workers will actually decrease welfare. Instead, once we adjust the share optimally, we can show that adding links increases welfare.

Theorem 1.5.3 *For every fixed $t \in (0, 1]$, $\phi(t, \mathcal{H}) \leq \phi(t, \mathcal{H} + ij)$. In particular, when $t = 1$, $\bar{w}(\mathcal{H}) \leq \bar{w}(\mathcal{H} + ij)$.*

Definition 7 *A hierarchy $\mathcal{H}$ is* **maximal** *if any member in N^k is linked to any member in N^{k+1}, for $k = 1, 2, \cdots, h - 1$.*

Adding links (with suitable adjustment of shares) means improving welfare, therefore we should link all the unlinked workers in adjacent levels (if the partition is fixed). In the end, we construct a maximal hierarchy, i.e, a multi-partition graph. Thus, we have the following corollary.

Corollary 1.5.4 *For hierarchies with N workers, the optimal hierarchy is a multi-partition graph if the number of workers on each level is fixed.*

For a maximal hierarchy, workers within each level are symmetric. As a special case, for leading by example $\mathcal{L}_{(1,N-1)}$, Hermalin (1998) shows that all $N - 1$ followers should get the same shares under optimal affine linear contracts. The following proposition shows that this holds for general maximal hierarchies.

Proposition 1.5.5 *For maximal hierarchies under optimal shares, members on the same level are assigned the same shares, hence work equally hard in equilibrium.*[12]

We still do not know exactly what the optimal shares are for any maximal hierarchy. Nevertheless, these results show that we can reduce the number of variables from N, the total number of workers, to h, the height of the hierarchy.

Starting from a multi-partition graph, we can split one level into multiple levels. Figure 1.7 explains such a procedure. We claim that this procedure also improves welfare after adjusting shares optimally.

Suppose $\mathcal{H}$ is a maximal hierarchy defined by the ordered partition $\mathcal{P} = \{N^1, \cdots, N^h\}$ of $\mathcal{N}$. Suppose $|N_k| \geq 2$ for some k. Pick $i \in N^k$, and let $\mathcal{H}'$ be the maximal hierarchy defined by the ordered partition $\mathcal{P}' = \{N^1, \cdots, N^{k-1}, \{i\}, N^k \backslash \{i\}, \cdots, N^h\}$.

[12] This result relies on $c(x) = \frac{1}{2}x^2$. It may not hold for other cost functions.

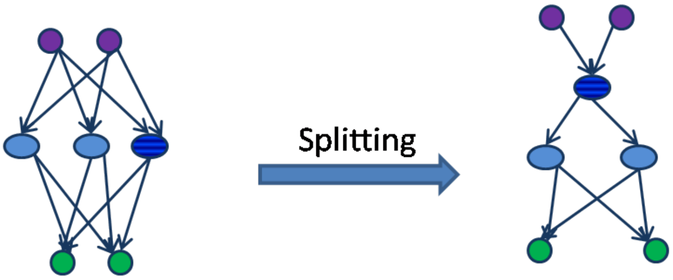


Figure 1.7: Splitting

Proposition 1.5.6 *For fixed $t \in (0,1]$, we have:*

$$\phi(t, \mathcal{H}) < \phi(t, \mathcal{H}')$$

In particular, for $t = 1$, we have $\bar{w}(\mathcal{H}) < \bar{w}(\mathcal{H}')$. The new hierarchy constructed by splitting is more efficient than the original one.

By repeated splitting until we have a partition that we cannot split at any level, eventually, we get the chain. Here is a map presenting the procedures to move from any hierarchy to the chain:

$$\text{hierarchy } \mathcal{H} \overset{\textit{add links}}{\Longrightarrow} \text{maximal hierarchy} \overset{\textit{splitting}}{\Longrightarrow} \textit{chain} \qquad (1.10)$$

Each step is welfare improving, which shows the optimality of the chain structure.

Theorem 1.5.7 *Among all hierarchies, the chain structure is optimal.*[13]

This theorem is the counterpart of Theorem 4.6 for simple hierarchies.

1.6 Extensions

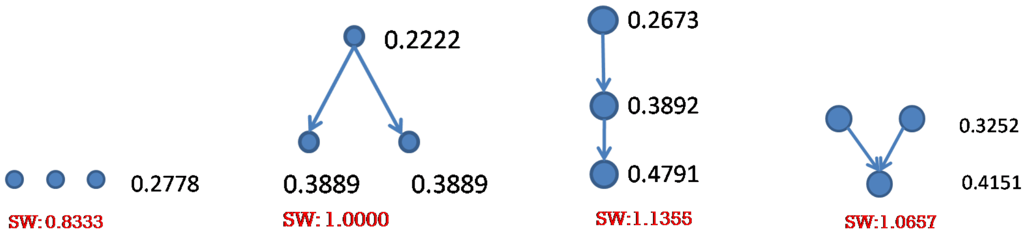
1.6.1 Who will be a leader?

Previously, we have only compared the aggregate welfare of the whole team across different hierarchies. In this section, we study an individual's incentive in various

[13] This result depends crucially on the fact that we are picking the equilibrium given in Theorem 1.5.1 for complicated hierarchies. Using other belief functions or picking other equilibria may reverse the ranking. See the discussion in the appendix about belief functions with the V structure. We do not have this problem when restricting attention to simple hierarchies.

hierarchies. As an example, with equal shares, leaders contribute more to the common project. Why should they do that? Who would ever want to a leader? To answer this question, first we review some results from section 2.

Figure 1.8 compares the equilibrium payoffs of three workers in various hierarchies with equal share $\mathbf{s}^{eq}$. The number to the right of each node is the payoff of that node when $\theta = 1$. SW denotes welfare when $\theta = 1$.



(a) Team $\mathcal{T}_3$ (b) Lead by Example Λ (c) Chain $\mathcal{C}_3$ (d) V structure

Figure 1.8: Individual payoffs with different hierarchies

A general fact illustrated in the figure is that a leader's equilibrium payoff is actually lower than her followers. This is consistent with Corollary 4.5. Another interesting observation is to compare the payoffs of the leader under the first three hierarchies. The leader's payoff in the Λ structure, 0.2222, is lower than he gets in the team structure, 0.2778. But the leader's payoff is higher in the chain, as the middle player is now contributing more to the project. Meanwhile, given the fact that a lump-sum transfer could be used to adjust individual payoffs without affecting the incentives, thus equilibrium efforts, of workers, it is reasonable to concentrate only on comparisons of aggregate welfare.

1.6.2 Limited Height

The previous analysis has shown that the chain is optimal among all hierarchies, but the height of the chain is too large to be realistic when N is large. Here we search for optimal hierarchies satisfying more realistic conditions, such as a constraint on the number of levels, being simple, and having a single leader. Formally, define

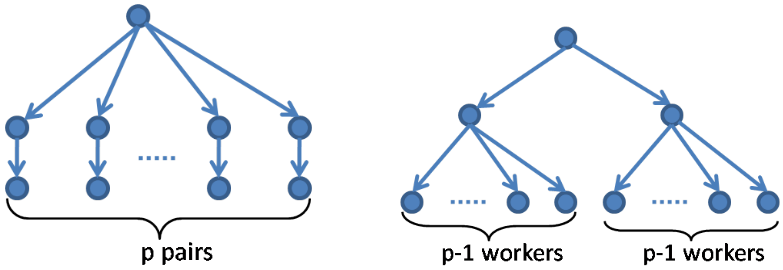
$$\mathcal{M}^s(N, K, 1) = \{\mathcal{H} |\ \mathcal{H} \text{ is simple, has height } K \text{ and one leader}\}.$$

Here we look for the optimal hierarchy in $\mathcal{M}^s(N, K, 1)$.[14]

[14] Equivalently, we could define the set $\mathcal{M}^s(N, K, 1)$ by including all the hierarchies with height at most K. The optimal element would be the same. The reason is that the constraint on height

For $K = 2$, the program is trivial, since the only hierarchy satisfying all of the three conditions is leading by example $\mathcal{L}_{(1,N-1)}$. The optimal sharing rule for this structure is completely solved in Hermalin (1998) for any N.

For $K = 3$, the problem gets tricky. For general $N \geq 3, K < N$, $\mathcal{M}^s(N, K, 1)$ is not a singleton. There could be a different number of middle managers, and different groups of followers for each middle manager.



(a) $\mathcal{H}^1$ (b) $\mathcal{H}^2$

Figure 1.9: two hierarchical structures in $\mathcal{M}^s(N, 3, 1)$

For example, if $N = 1+2p$ is odd and $p \geq 2$, Figure 1.9 presents two hierarchies $\mathcal{H}^1$ and $\mathcal{H}^2$ in $\mathcal{M}^s(N, 3, 1)$. For $\mathcal{H}^1$, there are p managers, and each has only one follower; for $\mathcal{H}^2$, there are only 2 managers, and each has p followers.

Proposition 1.6.1 *Hierarchy $\mathcal{H}^1$ is more efficient than $\mathcal{H}^2$:*

$$\bar{w}(\mathcal{H}^1) = \max_{s \in \Delta^N} w(\mathbf{s}, \mathcal{H}^1) \quad > \quad \max_{s \in \Delta^N} w(\mathbf{s}, \mathcal{H}^2) = \bar{w}(\mathcal{H}^2)$$

The intuition behind this proposition is as follows. Suppose share profile $\mathbf{s}$ is optimal for $\mathcal{H}^2$ and let l be the share of the leader, L. Then we can construct a profile $\mathbf{s}'$ for $\mathcal{H}^1$ as follows: L still gets l, while each manager gets $m = \frac{1}{3}\frac{1-l}{p}$ and every terminal worker gets $f = \frac{2}{3}\frac{1-l}{p}$. Note that $g(m, f) = g(\frac{1}{2}f, f) = f$, so all workers in $\mathcal{H}^1$ except L exert the same equilibrium effort. The sum of responsive coefficients for those $2p$ workers is

$$p\,(g(m, f) + f) = 2pf = \frac{4}{3}(1 - l)$$

We claim (with proof given in the appendix) that the sum of the responsive coefficients of all workers in $\mathcal{H}^2$ except the leader L under the contract $\mathbf{s}$ is less than

must bind for the optimal hierarchy in $\mathcal{M}^s(N, K, 1)$, since it must use the maximum height, i.e, K in this case.

$\frac{4}{3}(1-l)$, i.e,

$$\frac{4}{3}(1-l) \geq \sum_{j \neq L} k_j(\mathbf{s}, \mathcal{H}^2) \tag{1.11}$$

or

$$f = \frac{2}{3}\frac{1-l}{p} \geq \frac{\sum_{j \neq L} k_j(\mathbf{s}, \mathcal{H}^2)}{2p}.$$

Note $\eta(k) := k - \frac{1}{2}k^2$ is concave in k and increasing if $k < 1$, so by Jensen's inequality,

$$\begin{aligned} \sum_{j \neq L} \eta(k_j(\mathbf{s}, \mathcal{H}^2)) &= 2p \left(\sum_{j \neq L} \frac{1}{2p} \eta(k_j(\mathbf{s}, \mathcal{H}^2)) \right) \\ &\leq 2p \times \eta(\frac{\sum_{j \neq L} k_j(\mathbf{s}, \mathcal{H}^2)}{2p}) \leq 2p \times \eta(f) \end{aligned} \tag{1.12}$$

The last inequality holds because $\frac{\sum_{j \neq L} k_j(\mathbf{s}, \mathcal{H}^2)}{2p} \leq f = \frac{2}{3}\frac{1-l}{p} < \frac{2}{3}\frac{1}{p} < 1$. This shows the contribution to welfare by all workers except L is higher in $\mathcal{H}^1$ than in $\mathcal{H}^2$. The leader's incentive is higher in $\mathcal{H}^2$, because her share under the two cases is the same but the sum of the responsive coefficients of her followers is higher in $\mathcal{H}^2$ by equation 1.11. The old trick applies. We reduce the share of L by $\Delta > 0$ to make her incentive equal and then apply Theorem 4.9 to finish the proof.

The same argument can be used to show that hierarchy $\mathcal{H}^1$ is not only more efficient than $\mathcal{H}^2$, but also more efficient than any other hierarchy in $\mathcal{M}^s(N, 3, 1)$.

Proposition 1.6.2 *If $N = 1 + 2p$ is odd, $p \geq 2$, and $K = 3$, then hierarchy $\mathcal{H}^1$ is the most efficient in $\mathcal{M}^s(N, K, 1)$. That is,*

$$\bar{w}(\mathcal{H}^1) > \bar{w}(\mathcal{H}), \quad \forall \mathcal{H} \in \mathcal{M}^s(N, K, 1), \ \mathcal{H} \neq \mathcal{H}^1$$

1.6.3 Endogenous Information Acquisition

Previously, we have assumed that leaders are endowed with information, How do the leaders get the information in the first place? Presumably, the leaders exert costly research effort, such as sampling, running regressions, or consulting experts, to get more accurate information about the state. In this section, we study endogenous information acquisition in hierarchies.

For simplicity, assume $\mathcal{H}$ is simple, and there is a unique leader, L. She is the only one who will acquire information. To study this extension, insert one more stage between $t = 0$ (nature chooses θ) and $t = 1$ (the leader expends her effort x_L) in the game $G(\mathbf{s}, \mathcal{H})$:

- $t = 0.5$, the leader L exerts research effort $I \in \mathcal{I} = [I_0, I_1]$ and gets a signal s.

For each information structure I, $\theta^I \sim E^I\{\theta|s\}$ is the posterior. Let F^I be the C.D.F of θ^I. Furthermore we assume:

1. The support of θ^I is Θ for any $I \in \mathcal{I}$.
2. The utility of the leader is additively separable in both research effort and productive effort.
3. For $I < I'$, $\theta^{I'}$ is a mean-preserving spread (MPS) of θ^I.[15]

Condition 2 guarantees that the leader's research effort does not affect his signaling incentives. Risk-neutrality of workers and Condition 1 implies that the equilibrium characterization still applies, except that now we have to replace the state by the leader's point estimate in Theorem 4.3. Condition 3 means that the label of the information structure preserves the informativeness of the signal, that is, the higher is I, the more spread is the distribution of θ^I.

If research effort is verifiable (so it can serve as a contract contingency), then we only need to maximize expected social welfare of information minus the cost of research effort:

$$\mathcal{U}(\mathcal{H}) := \max_{I \in \mathcal{I}} \max_{\mathbf{s} \in \Delta^N} \left(\int_\Theta w(\mathbf{s}, \mathcal{H}) \theta^2 dF^I(\theta) - r(I) \right) \tag{1.13}$$

Here $r(I)$ is the cost of research effort I. Assume $r' > 0$, so a more accurate signal is more expensive. For convenience, define $v(I) := \int_\Theta \theta^2 dF^I(\theta)$ as the second moment of θ^I. Then condition 3 implies that v is monotone increasing in I. Rewrite equation (1.13) as:

$$\begin{aligned} \mathcal{U}(\mathcal{H}) &= \max_{I \in \mathcal{I}} \max_{\mathbf{s} \in \Delta^N} \left(w(\mathbf{s}, \mathcal{H}) v(I) - r(I) \right) \\ &= \max_{I \in \mathcal{I}} \left(\bar{w}(\mathcal{H}) v(I) - r(I) \right) \end{aligned}$$

Let $I^*(\mathcal{H})$ be the maximizer (assume it is unique, for simplicity). Then for two given hierarchies $\mathcal{H}$ and $\mathcal{H}'$, if $\bar{w}(\mathcal{H}') > \bar{w}(\mathcal{H})$, standard monotone comparative statics results (Milgrom and Shannon 1994) imply that $I^*(\mathcal{H}') \geq I^*(\mathcal{H})$. Greater marginal social value of information ($\bar{w}(\mathbf{s}, \mathcal{H})$) will induce higher research effort by the leader. Moreover, we have $\mathcal{U}(\mathcal{H}') \geq \mathcal{U}(\mathcal{H})$. As a corollary of Theorem 4.6, we can establish that the chain provides the greatest information acquisition and welfare.

[15] Each θ^I necessarily has the same mean by the law of iterated expectations: $E[\theta^I] = E[E[\theta|s]] = E[\theta], \forall I \in [I_0, I_1]$. See Rothschild and Stiglitz (1970) for the formal definition and properties of MPS.

Theorem 1.6.3 *Given assumptions 1-3 and verifiability of research effort, the chain $\mathcal{C}_N$ induces the highest research effort and yields the greatest expected welfare in the extended game with endogenous information acquisition.*

Now, we assume that we cannot write a contract contingent on the leader's research effort, either because research effort is not observable, or because it might be observable but hard to verify in court. Then the leader's research incentive comes from his private value of information, which, in general, is lower than the social value.

From Theorem 4.3, we know that the equilibrium payoff of the leader without information cost is

$$\pi_L(\mathbf{s},\mathcal{H}) = \theta^2 \left(s_L(\sum_{j\in N} k_j) - \frac{1}{2}k_L^2 \right) = \frac{1}{2}\theta^2 k_L(\mathbf{s},\mathcal{H})^2$$

which is monotonic in the leader's equilibrium responsive coefficient. Also, the leader's equilibrium payoff does not depend explicitly on other workers' efforts. The second equality uses the properties of g and Theorem 4.3.

The optimal contract is the solution to the following program:

$$\begin{aligned} \mathcal{U}^n(\mathcal{H}) : &= \max_{\mathbf{s}\in\Delta^N, I\in\mathcal{I}} (w(\mathbf{s},\mathcal{H})v(I) - r(I)) && (1.14)\\ \text{subject to:} \quad & \text{(IC-L)} \quad I \in \arg\max_{I'\in\mathcal{I}} \frac{1}{2}k_L(\mathbf{s},\mathcal{H})^2 v(I') - r(I') \end{aligned}$$

Let $\mathbf{s}^{n*}(\mathcal{H})$ and $I^{n*}(\mathcal{H})$ be the solution. IC-L is the incentive compatibility condition for the leader's research effort.

Theorem 1.6.4 *Assume conditions 1-3 and that research effort is not verifiable. Then the chain $\mathcal{C}_N$ is still the most efficient hierarchy, even if we take the leader's research incentive into account. That is,*

$$\mathcal{U}^n(\mathcal{C}_N) \geq \mathcal{U}^n(\mathcal{H})$$

for any simple hierarchy $\mathcal{H}$ with a single leader.

This result is quite intuitive. Since the leader's research incentive only depends on her responsive coefficient, for any $(\mathbf{s}, I)$ which satisfies IC-L under $\mathcal{H}$, we can find a new contract $\mathbf{s}'$ for $\mathcal{C}_N$ such that $k_L(\mathbf{s}',\mathcal{C}_N) = k_L(\mathbf{s},\mathcal{H})$, and $w(\mathbf{s}',\mathcal{C}_N) \geq w(\mathbf{s},\mathcal{H})$.[16] In particular, the leader's responsive coefficient is the same, so $(\mathbf{s}', I)$ satisfies IC-L under $\mathcal{C}_N$. Also, $w(\mathbf{s}',\mathcal{C}_N) \geq w(\mathbf{s},\mathcal{H})$, the marginal social value of

[16] See proofs of Lemma 4.8 and Theorem 4.6 for the construction of $\mathbf{s}'$.

information is also higher under the chain. Therefore, the chain both gives the leader higher incentive to acquire more accurate information and generates higher marginal social value of information. Both forces move in the same direction, so in the end, the chain wins.

Remark 3 *If research effort is verifiable, then there is no conflict between choosing optimal shares and incentivizing the leader for choosing the socially optimal research effort, as one can see from equation 1.13. When research effort is not verifiable, choosing optimal shares and incentivizing the leader for information acquisition are in conflict.*

In general, information is under-provided, i.e., $I^{n*}(\mathcal{H}) \leq I^*(\mathcal{H})$, because the marginal private benefit of information from the perspective of L is lower than the corresponding social value ($\pi_L(\mathbf{s}, \mathcal{H}) < w(\mathbf{s}, \mathcal{H})\theta^2 \leq \bar{w}(\mathcal{H})\theta^2$). Of course, the maximal obtainable welfare is lower if research effort is not verifiable, i.e., $\mathcal{U}^n(\mathcal{H}) \leq \mathcal{U}(\mathcal{H})$. Moreover, in general $\mathbf{s}^{*n}(\mathcal{H}) \neq s^*(\mathcal{H})$, so we should modify the shares $s^*(\mathcal{H})$ to give the leader enough incentive for research. Nevertheless, the chain is the best given all these inefficiencies. Other hierarchies perform even worse.

1.6.4 Applications in fund-raising

A natural application of the model is charity fund-raising, which is similar to public good provision. Vesterlund (2003) and Andreoni (2006) emphasize the importance of leadership giving in charitable fund-raising, which serves as a signal to other givers that the charity is of high quality. Using our terminology, they show the superiority of sequential fund-raising, which corresponds to $\mathcal{L}_{(1,N-1)}$, over simultaneous fund-raising, which corresponds to $\mathcal{T}_N$.[17]

Given the optimality of the chain structure, a charity could raise more money by implementing the chain $\mathcal{C}_N$, i.e., placing potential donors in a line and asking them to donate one after the other. In particular, the charity should not reveal the entire donation history to future givers. A drawback to the chain is that it requires more steps to complete the fund-raising. If delay is costly to the charity, the techniques and results of this paper might still be useful for suggesting better ways to organize the fund-raising campaign. Rather than having each donor on a separate tier, donors could be organized into subgroups, with the total donation of each subgroup revealed only to the next tier. Smaller subgroups allow for a larger number of tiers, thus raising more money, but also results in a longer delay. The optimal configuration involves a trade-off between the costs of delaying and the

[17] For experimental evidence, see Protter *et al.*, (2001,2005).

benefits of the funds. Carefully designed future experiments should be able to test this theory in the field.

1.7 Conclusion

This paper highlights the importance of hierarchical structures from the perspective of information flow and signaling effects associated with dissemination of information. In a team production framework, we show the optimality of the chain structure from three perspectives: maximizing dynamic signaling effects, motivating efforts of all members, and providing strong incentives for the leader's information acquisition.

This paper isolates one feature, signaling effects, of organization design. In reality, there are other forces, such as communication, adaptation and coordination, which are also relevant for the design of organizations. Also, I model leaders as the source of information. There are many other features of leadership which I have not addressed. For example, Bolton *et al.* (2008) show that a resolute leader can achieve a better outcome for an organization faced with conflict between adaptation and coordination as a resolute leader overestimates the precision of her prior belief and hence is less responsive to new information. They show that the coordination benefit from a resolute leader generally outweighs the cost of mal-adaptation.[18] Adding these components into this model may balance the signaling effect which is dominant in this paper, thus lead to more realistic predictions about optimal hierarchies. Also, there might be other transaction costs, such as delaying, or communication costs, associated with each hierarchy. A related question is: Does the optimal hierarchy get longer and thinner, or the opposite, as transaction costs drop? Adding these costs will shed some light on our understanding of real organizational design problems. A detailed analysis of these new features requires another paper, and I plan to address these issues in the future.

[18] See Bolton *et al.* (2010) for a survey of key elements of leadership.

Chapter 2

Picking Winners in Rounds of Elimination

This chapter is based on joint work with Suzanne Scotchmer.

2.1 Introduction

We study economic environments where a principal must select projects or agents from a pool, but cannot observe a candidate's ability or the project's intrinsic worth. A technique for solving this problem is to cast a wide net, and then to eliminate agents or projects that do not perform well. For example, this is how professors are hired. A department initially hires a large pool of assistant professors, gives them a few years to demonstrate their worth, and then makes an up-or-out evaluation. At the full professor stage, the survivors are evaluated again. Those who are not promoted typically leave.

There are many other selection arenas that use a similar technique. Venture capitalists may give early funding to many young start-up firms, but cut them loose ruthlessly when they fail to perform. Drug testing is similar. After the first round of testing, many drugs are dropped, while others go on to another round.

In this paper we ask how such rounds of elimination should be structured. Should there be many rounds or only a few? How should the structure depend on cost? Should standards become tougher or more lax in later rounds? How should the selection at a later round incorporate information generated at earlier rounds?

Different versions of this problem call for different stylizations. In the stylization here, there are natural periods of time in which agents or projects generate signals. The arrival of independent signals in successive periods leads naturally to rounds of elimination. This structure arises naturally in academic life, where evaluation periods are established by policy. Hiring contests are also structured this way. It is common for a software engineering firm to test its applicants by giving them problems to solve. Candidates must pass all the rounds of elimination in order to be considered. How should the rounds of elimination and the use of the signal be structured?

To isolate the issues, we consider only two periods, with two signals (x_1, x_2), independently drawn from a distribution determined by an unobservable ability parameter, μ. The principal wants to select for high values of μ. He can select in a single round of elimination, waiting until the end of period 2 and using both signals, or he can select in double rounds of elimination, using x_1 to winnow the candidates at round one, and then using x_2 to winnow them further. We assume that in both schemes, he is constrained to end up with the same number of survivors.

We distinguish between selection schemes with memory and those without, as in Scotchmer (2008). In a selection scheme with memory, the selection at round two can use the signal generated at round one as well as the signal generated at round two. In a selection scheme without memory, selection at each round can only depend on the signal generated in that round. Sports tournaments are typically selection schemes without memory, whereas promotion in the academic hierarchy typically has memory.

Some of our conclusions are obvious, or at least very intuitive, once stated. However, even the "obvious" conclusions are not always true. They require conditions on the distributions, which we illuminate in this paper. We give a general characterization of optimal selection sets, but also illuminate their special structure when probability densities are log-supermodular. Our work is related to an earlier literature on optimization that uses supermodularity for the conclusion that optimal control variables move monotonically with underlying parameters (Topkis, 1978, Milgrom and Weber, 1982, Milgrom and Shannon, 1995, Athey, 2002). We use log- supermodularity for a different purpose. Instead of using log-supermodularity to characterize how control variables move with parameters, we use log-supermodularity to characterize optimal selection sets.

Our main conclusions are:

- When the joint distribution of μ and x satisfy log-supermodularity, optimal selection sets can be characterized by threshold values on the signals.

- Conditional on a given number of ultimate survivors, the average ability of survivors in a single round of elimination is higher than in any double round of elimination, and the average ability of survivors is higher if the selection scheme has memory than if not.

- A higher cost of holding on to candidates should lead to more stringent screening at round one, less stringent screening at round two, and lower average ability among ultimate survivors.

- In double rounds of elimination without memory (depending on a hazard rate condition), the selection standard should be tougher in round one than in round two.

- If it is optimal to use a sufficient statistic for selection in a single round of elimination, then it is optimal to use the same sufficient statistic in the ultimate round of double elimination, even though the sample has been winnowed in round one, using partial information.

The last point, which follows from the factorization theorem, is perhaps the least intuitive. One might have thought that, in double rounds of elimination, extra weight should be given to x_2 at the second round. The signal x_1 was already used for selection at the first round, and conditional on survival at round one, the signal x_1 is likely to exhibit some "good luck bias." Nevertheless, the same sufficient statistic should be used at the second round as if no prior selection had taken place, although with a less stringent screening standard. No extra weight should be given to x_2, even if there has already been screening on x_1.

In section 2, we describe a simple model. In section 3 we characterize how the selection set should be chosen for a single round of elimination, and record some well known features of probability distributions that lead to monotonic selection criteria. We also give examples, showing that the parameter of interest, μ, may be interpreted in many ways, such as the mean of a distribution, a measure of upside or downside risk, or expected waiting time for an arrival. In section 4 we discuss double rounds of elimination with memory, and in section 5 we discuss double rounds of elimination without memory.

Since we conclude that all the information should be used at every round, section 5 is mainly of interest because many selection schemes ignore or de-emphasize information from earlier rounds. Our work implies that, to explain this, one must look elsewhere than simple screening. For example, moral hazard could be a justification. De-emphasizing earlier success maintains an incentive to work harder in later rounds. We intentionally put aside moral hazard problems, because our objective is to isolate the screening problem. Elements of our characterizations will remain when screening and moral hazard are combined.

2.2 The model

We assume that agents (or projects) are endowed with an underlying parameter μ, which is unobservable. The value of μ is therefore a random variable from the perspective of an observer. The observer wants to select a given number of agents or projects in a way that maximizes the expected value of μ. The underlying parameter μ could be, for example, the profitability of a project, the potential market size of a new innovation, the ability of an assistant professor, the assistant professor's upside potential, or the rate at which the assistant professor thinks of good ideas. The prior distribution on μ is given by a density function h.

The objective of the selection scheme is to maximize the expected value of μ among survivors, but μ does not need to be the mean of the distribution. As long as the log-supermodularity condition below is satisfied, our optimal selection theorems apply. In section 3.2 below we give examples of how μ might be interpreted.

The agents (or projects) generate signals $x \in \mathbf{R}^2$ (more generally, $x \in \mathbf{R}^n$) where the draws are assumed to be independent conditional on the underlying value of μ. Each x_i has probability distribution $F(\cdot, \mu)$. Throughout we maintain the assumption that the distribution of signals is atomless with density $f(\cdot, \mu)$.

This simple structure with $x \in \mathbf{R}^2$ permits two rounds of elimination. Agents or projects might be eliminated at the first round, based on x_1, or at the second round, based on (x_1, x_2). A *single round of elimination* means that all agents are kept in the pool until the end of the second period, and selection uses the information generated in both periods, (x_1, x_2). *Double rounds of elimination* mean that some of the agents are eliminated after round one, using only the information x_1. At the second round, selection can take place using both (x_1, x_2) or only x_2. This is the distinction between rounds of elimination *with memory* and *without memory.*

We will use the following notation when it is convenient and not ambiguous:[1] When we write an integral sign without delimiters, we mean the integral on the full support.

$$\begin{array}{rcl} p(x_1, x_2, \mu) & = & f(x_1, \mu) f(x_2, \mu) h(\mu) \\ p(x_1, x_2) & = & \int p(x_1, x_2, \mu)\, d\mu \\ p(x_1) & = & \int \int p(x_1, x_2, \mu)\, dx_2 d\mu = \int f(x_1, \mu) h(\mu)\, d\mu \\ p(x_1, \Delta_2) & = & \int_{\Delta_2} p(x_1, x_2)\, dx_2 \\ p(\Delta_1, x_2) & = & \int_{\Delta_1} p(x_1, x_2)\, dx_1 \end{array}$$

Conditional probabilities will also be expressed in this notation:

$$\begin{array}{rcl} p(\mu | x_1, x_2) & = & p(x_1, x_2, \mu) / p(x_1, x_2) \\ p(x_2 | x_1) & = & p(x_1, x_2) / p(x_1) \\ p(\Delta_2 | x_1) & = & p(x_1, \Delta_2) / p(x_1) \\ p(x_2 | x_1, \Delta_2) & = & p(x_2 | x_1) / p(\Delta_2 | x_1) \\ p(x_1 | \Delta_1, x_2) & = & p(x_1 | x_2) / p(\Delta_1 | x_2) \end{array}$$

[1] We abuse notation to avoid awkwardness, hopefully without confusion. Instead of writing, for example, $p_{(\mu, X_1, X_2)}(\cdot)$ and $p_{(X_1|X_2)}(\cdot)$, as the names of the density functions, we simply write $p(\mu, x_1, x_2)$ and $p(x_1|x_2)$. That is, we write the same thing to refer to the functions themselves as well as to point values of the functions. The context will indicate which interpretation is appropriate.

We will also use sufficient statistics. Using the factorization theorem, a function $\sigma : \mathbf{R}^2 \rightarrow \mathbf{R}$ of the signals is *sufficient for* μ if the joint density $p\left(x_1, x_2, \mu\right)$ can be written as

$$p\left(x_1, x_2, \mu\right) = q(x_1, x_2)\theta(\sigma(x_1, x_2), \mu) \tag{2.1}$$

for functions $q, \sigma : \mathbf{R}^2 \rightarrow \mathbf{R}$ and $\theta : \mathbf{R}^2 \rightarrow \mathbf{R}$.

2.3 Single Round of Elimination

We first consider a single round of elimination, using the information generated in both periods.

In a single round of elimination, the *selection set* is a subset Δ of $\mathbf{R}^2$ such that only the agents or projects that generate signals $(x_1, x_2) \in \Delta$ are chosen. Others are thrown away. When integrated over a set of signals $\Delta \in \mathbf{R}^2$, the total number of survivors is

$$S^s\left(\Delta\right) =: \int_{\Delta} \int p(x_1, x_2, \mu) d\mu dx_1 dx_2$$

and their total ability is

$$V^s\left(\Delta\right) =: \int_{\Delta} \int \mu p(x_1, x_2, \mu) d\mu dx_1 dx_2$$

We pose the optimization problem with a constraint on the probability of survival (or number of survivors), $S^s\left(\Delta\right) = \Lambda$, $\Lambda \in (0, 1)$. Reducing the set Δ leads to fewer survivors. Our problem is to choose the selection set such that the expected ability of survivors is maximized.

The problem we wish to solve is

$$\max_{\Delta}\ V^s\left(\Delta\right) \text{ subject to } S^s\left(\Delta\right) = \Lambda \tag{2.2}$$

Following is a general characterization of the solution.

Theorem 2.3.1 (Single round of elimination: the optimal selection set) *There exists a finite number α such that, for every solution Δ to (2.2),*

$$\begin{aligned} E\left(\mu|x_1, x_2\right) &\geq \alpha \quad \textit{for a.e.} \quad x \in \Delta \\ E\left(\mu|x_1, x_2\right) &\leq \alpha \quad \textit{for a.e.} \quad x \in \mathbf{R}^2\backslash\Delta \end{aligned} \tag{2.3}$$

If $\hat{\Delta}$ and Δ are two solutions to (2.2), then

$$E\left(\mu|x_1, x_2\right) = \alpha \textit{ for a.e. } x \in (\hat{\Delta}\backslash\Delta) \cup (\Delta\backslash\hat{\Delta}) \tag{2.4}$$

Proof: It will be convenient to state the optimization problem using a function g defined as the expected value of μ, given (x_1, x_2).

$$g(x_1, x_2) = \int \mu \, \frac{p(x_1, x_2, \mu)}{\int p(x_1, x_2, \mu)\, d\mu} d\mu = E(\mu | x_1, x_2)$$

Let $d\nu = \left[\int p(x_1, x_2, \mu)\, d\mu\right] dA$ on the interior of its support in $\mathbf{R}^2$. Because $d\nu$ and dA are absolutely continuous with respect to each other, a set that has measure zero with respect to Lebesgue measure also has measure zero with respect to ν measure.

The problem in a single round of elimination can be stated as

$$\max_{\Delta} \int_{\Delta} g \, d\nu \quad \text{subject to:} \quad \int_{\Delta} 1 \cdot d\nu = \nu(\Delta) = \Lambda \tag{2.5}$$

In appendix A,we first show that for each Λ, there exists a set Δ with the property that:

1. $\nu(\Delta) = \Lambda$,
2. There exists a finite number α such that $g \geq \alpha$ on Δ, and $g \leq \alpha$ on the complement.

We now show that this set is optimal. In particular, for any set A with $\nu(A) = \Lambda$, it must be the case that:

$$\int_A g d\nu \leq \int_{\Delta} g d\nu$$

The first observation is that:

$$\int_{A \backslash \Delta} g d\nu \leq \int_{A \backslash \Delta} \alpha d\nu = \nu(A \backslash \Delta)\alpha = \nu(\Delta \backslash A)\alpha = \int_{\Delta \backslash A} \alpha d\nu \leq \int_{\Delta \backslash A} g d\nu$$

These inequalities hold because $g \leq \alpha$ on $A \backslash \Delta \subset \mathbf{R}^2 \backslash \Delta$, $\alpha \leq g$ on $\Delta \backslash A \subset \Delta$, and $\nu(\Delta \backslash A) = \nu(\Delta) - \nu(A \cap \Delta) = \nu(A) - \nu(A \cap \Delta) = \nu(A \backslash \Delta)$. Thus,

$$\int_A g d\nu = \int_{A \cap \Delta} g d\nu + \int_{A \backslash \Delta} g d\nu \leq \int_{A \cap \Delta} g d\nu + \int_{\Delta \backslash A} g d\nu = \int_{\Delta} g d\nu$$

This shows (2.3).

For (2.4), suppose that $\hat{\Delta}$ is another optimal solution. Then $\int_{\hat{\Delta} \backslash \Delta} g d\nu = \int_{\Delta \backslash \hat{\Delta}} g d\nu$. For almost every $x \in \hat{\Delta} \backslash \Delta$, $g \leq \alpha$ because $x \notin \Delta$, but $g \geq \alpha$ because

$x \in \hat{\Delta}$. Therefore $g = \alpha$, and similarly for $\Delta \backslash \hat{\Delta}$. Hence, $g = \alpha$ on $(\hat{\Delta} \backslash \Delta) \cup (\Delta \backslash \hat{\Delta})$, except on a set of measure zero. ■

Theorem 2.3.1 implies that a solution Δ is coupled with a value α that represents the expected ability of the marginal agent. However, without further assumptions, the value α that accompanies Δ is not necessarily unique. When α is not unique, the largest such value is of particular interest, because it represents the infimum of $E(\mu|x_1, x_2)$ on subsets of selected signals (agents) that have positive measure.

For any solution Δ coupled with a particular α, we can get the other solutions by replacing the part of Δ where $E(\mu|x_1, x_2) = \alpha$ with another set of the same measure where $E(\mu|x_1, x_2) = \alpha$. However, when $\nu(x \in \mathbf{R}^2 : E(\mu|x) = \alpha) = 0$, the optimal selection set is almost unique. By this we mean that if Δ is a solution coupled with α and and $\hat{\Delta}$ is also a solution, it is coupled with the same α, and either that $\nu((\hat{\Delta} \backslash \Delta) \cup (\Delta \backslash \hat{\Delta})) = 0$ or $\nu(\Delta \backslash \hat{\Delta})) = \nu(\hat{\Delta} \backslash \Delta)) = 0$. Every optimal solution is just Δ except on a zero measure set,which is impossible to detect using integration.

When there is a sufficient statistic for μ, Theorem 2.3.1 can be restated using the sufficient statistic. For each value $\bar{\sigma} \in \mathbf{R}$, define

$$\bar{E}(\mu|\bar{\sigma}) = \int \mu \frac{\theta(\bar{\sigma}, \mu)}{\int \theta(\bar{\sigma}, \mu) d\mu} d\mu$$

Then it is easy to show that $E(\mu|x_1, x_2)$ has the same value for every signal (x_1, x_2) in the set $\{(x_1, x_2) | \sigma(x_1, x_2) = \bar{\sigma}\}$, and

$$\bar{E}(\mu|\sigma(x_1, x_2)) = E(\mu|x_1, x_2) \tag{2.6}$$

Theorem 2.3.1 thus implies

Corollary 2.3.2 *Suppose that σ is a sufficient statistic for μ. There exists a finite number α such that, for every solution Δ to (2.2),*

$$\bar{E}(\mu|\sigma(x_1, x_2)) \geq \alpha \quad \textit{for a.e.} \quad x \in \Delta$$

$$\bar{E}(\mu|\sigma(x_1, x_2)) \leq \alpha \quad \textit{for a.e.} \quad x \in \mathbf{R}^2 \backslash \Delta$$

If $\hat{\Delta}$ and Δ are two solutions to (2.2), then

$$\bar{E}(\mu|\sigma(x_1, x_2)) = \alpha \textit{ for a.e. } x \in (\hat{\Delta} \backslash \Delta) \cup (\Delta \backslash \hat{\Delta}) \tag{2.7}$$

2.3.1 Single Round: Promotion thresholds and monotonicity

The characterization in Theorem 2.3.1 and Corollary 1 is too general to be useful. For example, it does not say that if the signal $(\hat{x}_1, \hat{x}_2)$ is larger than some signal (x_1, x_2) in the selection set, then the larger signal $(\hat{x}_1, \hat{x}_2)$ should also be selected. And it does not say that if the value of the sufficient statistic $\sigma(\hat{x}_1, \hat{x}_2)$ is larger than a value $\sigma(x_1, x_2)$ that would be selected, then the larger value of the sufficient statistic should also be selected. These are intuitive conclusions; if they do not hold, then the signal (x_1, x_2) has no natural interpretation.

We will use the mathematical structure of supermodularity. This structure is used widely in economics, following Topkis (1978, p. 310), Milgrom and Weber (1982), Milgrom and Shannon (1995), and Athey (1982). The literature is concerned with monotone comparative statics, that is, monotonicity of optimal choices with respect to underlying variables. If an optimand is supermodular with respect to the appropriate variables, then the optimizer is a monotonic function of the underlying parameters. Our application here is concerned with optimal selection sets for random variables rather than with optimal control variables. Monotonicity leads to the conclusion (among others) that selection sets can be characterized by threshold values.

For convenience, we state the monotonicity assumptions on the density function p. However, they follow from the same properties of f.

The density p satisfies the *monotone likelihood ratio property if*

$$\begin{aligned}
\frac{p(x_1', x_2, \mu')}{p(x_1, x_2, \mu')} &\geq \frac{p(x_1', x_2, \mu)}{p(x_1, x_2, \mu)} \text{ whenever } x_1' > x_1, \mu' > \mu \\
\frac{p(x_1', x_2', \mu)}{p(x_1', x_2, \mu)} &\geq \frac{p(x_1, x_2', \mu)}{p(x_1, x_2, \mu)} \text{ whenever } x_1' > x_1, x_2' > x_2
\end{aligned}$$

and the equivalent statements hold when x_1 and x_2 are reversed.

Assuming that p is twice differentiable, we will say that the density p is *log supermodular (strictly log supermodular)* if the cross partial derivative of $\log p$ in any of $(x_1, \mu), (x_2, \mu), (x_1, x_2)$ is nonnegative (positive).This is not the general definition, but is equivalent to the general definition when p is differentiable Topkis (1978, p.310). See the papers cited above for the definition and underlying mathematical structure. We will also refer to log supermodularity of θ in (2.1).

We will also use first-order stochastic dominance. Let two distributions F and G have common supports in $\mathbf{R}$. We say that F *first-order stochastically dominates* G if $F(t) \leq G(t)$ for all t in the supports.

An important fact is that, if F first order dominates G, the expected value of the random variable, or an increasing function of the random variable, is larger when distributed as F than when distributed as G. We will use a slight extension of this fact, stated in the following lemma.

Lemma 2.3.3 *Let F and G be two distributions on supports contained in $\mathbf{R}$ such that F first-order dominates G. Suppose that two functions $u, v : \mathbf{R} \to \mathbf{R}$ have the properties that (1) both are nondecreasing and (2) $u(t) \geq v(t)$ for all t. Then:*

$$\int u(t)dF(t) \geq \int v(t)dG(t)$$

Proof:

$$\begin{aligned} & \int u(t)dF(t) - \int v(t)dG(t) \\ = & \int \underbrace{(u(t) - v(t))}_{\geq 0} dF(t) + \left\{ \int v(t)dF(t) - \int v(t)dG(t) \right\} \geq 0 \end{aligned}$$

The first term is nonnegative because the integrand is nonnegative. Nonnegativity of the second term follows because F first-order dominates G and because v is nondecreasing. ■

The following remark records some relationships among the definitions that are used heavily below. The first two bullet points reflect the fact that supermodularity is preserved by integration. This property underlies the analysis of Athey (2002), who extended monotone comparative statics to problems where the optimand is the expected value of a supermodular function.

Remark 4 *Suppose that $p(x_1, x_2, \mu)$ is log supermodular. Then*

- *Any marginal density function derived from p is log supermodular. For example, $\int p(x_1, x_2, \mu)\, d\mu$ is log supermodular in (x_1, x_2).*
- *Any conditional density function derived from p, such as $p(x_1, \mu | x_2)$, is log supermodular.*
- *Any conditional density function derived from p satisfies the monotone likelihood ratio property.*
- *If $x_1 > \hat{x}_1$, $p(\mu, x_2 | x_1)$ stochastically dominates $p(\mu, x_2 | \hat{x}_1)$ (and symmetrically for the other conditional distributions).*
- *The expected value of μ, conditional on x_1, is increasing in x_1 (symmetrically, x_2).*

When we assume that p is log supermodular below, we are adopting all the properties in the Remark.

We also use the following lemma, which links monotonicity properties of a sufficient statistic to log supermodularity of the underlying density function.

Lemma 2.3.4 *If p is strictly log supermodular and twice differentiable, and can be written as (2.1), then σ and θ can be chosen such that θ is strictly log supermodular and σ is increasing in its arguments.*

Proof of Lemma: We want to show that, as long as there is any pair of functions σ and θ such that (2.1) holds, then there is a pair of function σ and θ such that $\frac{\partial^2}{\partial\sigma\partial\mu}\log\theta\left(\sigma,\mu\right)>0$. Write, for $i=1,2$,

$$\begin{aligned}\frac{\partial^2}{\partial x_1\partial\mu}\log p\left(x_1,x_2,\mu\right) &= \frac{\partial^2}{\partial x_1\partial\mu}\log g\left(x_1,x_2\right)+\frac{\partial^2}{\partial\sigma\partial\mu}\log\theta\left(\sigma\left(x_1,x_2\right),\mu\right)\frac{\partial}{\partial x_1}\sigma\left(x_1,x_2\right)\\ &= \frac{\partial^2}{\partial\sigma\partial\mu}\log\theta\left(\sigma\left(x_1,x_2\right),\mu\right)\frac{\partial}{\partial x_1}\sigma\left(x_1,x_2\right)\\ \frac{\partial^2}{\partial x_2\partial\mu}\log p\left(x_1,x_2,\mu\right) &= \frac{\partial^2}{\partial\sigma\partial\mu}\log\theta\left(\sigma\left(x_1,x_2\right),\mu\right)\frac{\partial}{\partial x_2}\sigma\left(x_1,x_2\right)\end{aligned}$$

Because $\frac{\partial^2}{\partial x_i\partial\mu}\log p\left(x_1,x_2,\mu\right)>0$, $\frac{\partial}{\partial x_i}\sigma\left(x_1,x_2\right)\neq 0$ for $i=1,2$. If $\frac{\partial}{\partial x_1}\sigma\left(x_1,x_2\right)>0$, then $\frac{\partial^2}{\partial\sigma\partial\mu}\log\theta\left(\sigma\left(x_1,x_2\right),\mu\right)>0$, as required, and $\frac{\partial}{\partial x_2}\sigma\left(x_1,x_2\right)>0$. If $\frac{\partial}{\partial x_1}\sigma\left(x_1,x_2\right)<0$, then $\frac{\partial^2}{\partial\sigma\partial\mu}\log\theta\left(\sigma\left(x_1,x_2\right),\mu\right)<0$ and $\frac{\partial}{\partial x_2}\sigma\left(x_1,x_2\right)<0$. In that case, define $\tilde{\sigma}=-\sigma$. Then $\tilde{\sigma}$ is also sufficient for μ, and $\theta\left(\tilde{\sigma},\mu\right)$ is log supermodular. ■

The following theorem shows that, if there is a sufficient statistic, the optimal selection set can be written as a threshold value on that statistic, that is, $\left\{\left(x_1,x_2\right):\sigma\left(x_1,x_2\right)\geq k\right\}$. If the selection scheme has memory, the optimal selection set cannot generally be written, for example, as $\left\{\left(x_1,x_2\right):x_2\geq k_1,x_2\geq k_2\right\}$.

Theorem 2.3.5 (Single round of elimination and a sufficient statistic) *Suppose that p is log supermodular, and let Δ be a selection set that solves* (2.2)*. Suppose that σ is a sufficient statistic for μ. Then there exists $\bar{\sigma}$ such that the optimal selection set can be written as*

$$\Delta=\left\{\left(x_1,x_2\right):\sigma\left(x_1,x_2\right)\geq\bar{\sigma}\right\}$$

Proof: This follows from Theorem 2.3.1, and from (2.6). With log supermodularity, σ is increasing in its arguments, as is $\bar{E}\left(\mu|\bar{\sigma}\right)$. ■

In the next section, we give some examples to show the different meanings that the parameter μ can take. The most familiar case is where μ is the distribution mean, and the sufficient statistic is the mean of the sample. We also consider examples where μ is an extreme point of the support, interpreted as a measure of upside risk or downside risk.

2.3.2 Examples

Normal Distribution. Let x be a single random draw from a normal distribution with unknown mean μ and known variance v. Then the distribution of x conditional on μ is

$$f(x,\mu) = \frac{1}{\sqrt{2\pi v^2}} e^{-\frac{(x-\mu)^2}{2v^2}},\ x \in \mathbf{R}$$

One can check directly that f is log supermodular. When a sample $(x_1, x_2, ..., x_n)$ is available, the joint distribution conditional on μ is

$$\begin{aligned}\Pi_{i=1}^{n} f(x_i,\mu) &= \left(\frac{1}{2\pi v^2}\right)^{n/2} e^{-\frac{\sum(x_i-\mu)^2}{2v^2}} = \left(\frac{1}{2\pi v^2}\right)^{n/2} \exp\left\{-\frac{\sum {x_i}^2 - 2\mu \sum x_i + \mu^2}{2v^2}\right\} \\ &= \left(\frac{1}{2\pi v^2}\right)^{n/2} \exp\left\{-\frac{\sum {x_i}^2}{2v^2}\right\} \exp\left\{\frac{2\mu n \bar{x}}{2v^2}\right\} \exp\left\{-\frac{\mu^2}{2v^2}\right\}\end{aligned}$$

Using the factorization theorem, the sample mean, $\bar{x}$, is a sufficient statistic for μ.

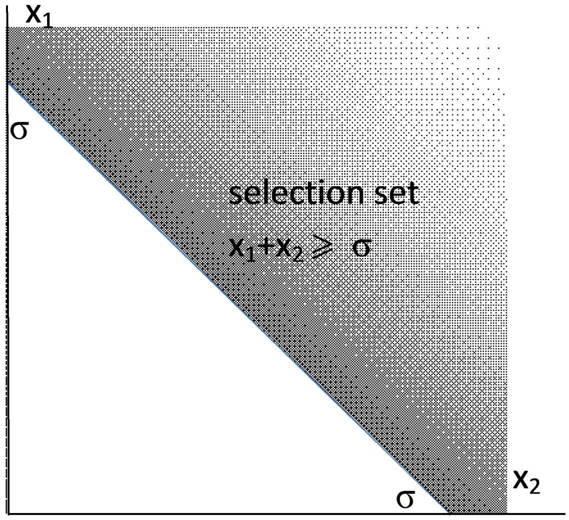


Figure 2.1: selection on the mean

Applying Theorem 2.3.5, the lower bound of the optimal selection set is an affine line with slope -1, that is $\Delta = \{(x_1,x_2)|x_1 + x_2 \geq \sigma\}$ for an appropriate value of the sufficient statistic, σ. This is shown in figure 2.1.

General exponential distributions. In the exponential family, the distribution of a single random draw, x, conditional on a parameter μ, can be expressed as

$$f(x,\mu) = h(x)\exp(\eta(\mu)\sigma(x) \ - \ A(\mu)),\ x \in \mathbf{R}$$

Then, using the factorization theorem for sufficient statistics, σ is a sufficient statistic for μ. Provided $\eta'(\mu)\sigma'(x) > 0$ (both are increasing or both are decreasing), the density function is log supermodular.

When there are two random draws, the joint density is $f(x_1,\mu) f(x_2,\mu)$, and the sufficient statistic is $\sigma(x_1) + \sigma(x_2)$.

Waiting Time: The random variable x with the following density represents a waiting time, conditional on μ :

$$f(x,\mu) = \frac{e^{-x/\mu}}{\mu}, \mu > 0, x \in \mathbf{R}_+$$

The waiting time itself, x, is a sufficient statistic. If two waiting times are measured, the sufficient statistic is their sum.

Maximizing the upside risk. Suppose x_1 and x_2 are independent and uniformly distributed on the interval $[0,\mu]$. The density of each random draw x is

$$f(x,\mu) = \frac{1}{\mu}1_{\{0\leq x\leq\mu\}}$$

where $1_{\{0\leq x\leq\mu\}}$ is the indicator function on the set $[0,1]$. This density function is log supermodular in (x,μ), because both $1_{\{0\leq x\leq\mu\}}$ [2] and $\frac{1}{\mu}$ are logsupermodular in (x,μ), and the product of nonnegative log supmodular functions is log supermodular. Therefore $E(\mu|x_1,x_2)$ is weakly increasing in x_1 and x_2.

Let

$$\sigma(x_1,x_2) = \max(x_1,x_2)$$

The probability density of (x_1,x_2) can be written as follows:

$$f(x_1,\mu)f(x_2,\mu) = \frac{1}{\mu}1_{\{0\leq x_1\leq\mu\}}\frac{1}{\mu}1_{\{0\leq x_2\leq\mu\}} = \left(1_{\{0\leq\min(x_1,x_2)\}}\right)\left(\frac{1}{\mu^2}1_{\{\max(x_1,x_2)\leq\mu\}}\right)$$

Therefore, using the factorization theorem, σ is sufficient for μ. Further, there is an unbiased estimator of μ, β_1, that increases with σ. For this estimator,

$$E(\mu|x_1,x_2) = \beta_1(\max(x_1,x_2))$$

[2] This can be checked directly from the general definitions, which we have not reprised here. For our purpsoes it is enough to cite Lemma 3 of Athey (2002), which tells us that the indicator function with values $1_{A(\mu)}(x)$ is logsupermodular in (x,μ) if and only if the set $A(\mu)$ is a sublattice. The condition holds because $A(\mu)$ is the interval $[0,\mu]$.

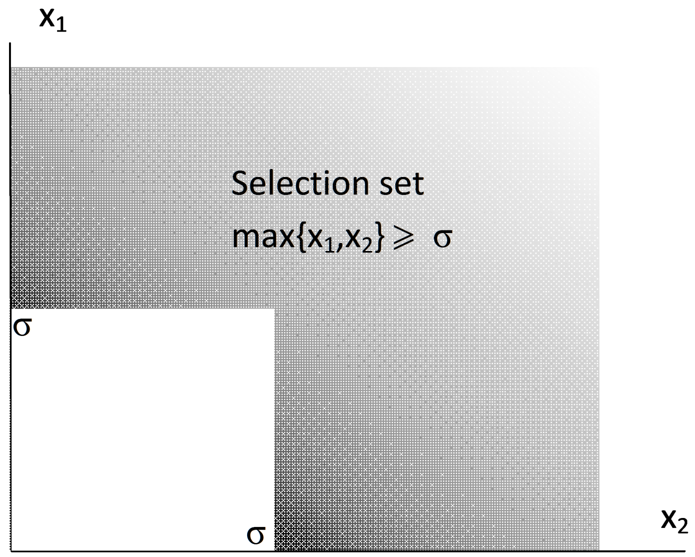


Figure 2.2: Selection for maximizing upside risk

For some number α, an upper contour set of $E(\mu|x_1,x_2)$, hence the selection set Δ, takes the following form:

$$\Delta = \{(x_1,x_2)|\sigma(x_1,x_2) \geq \alpha\} = \{(x_1,x_2)|\max(x_1,x_2) \geq \alpha\}$$

Minimizing the downside risk. Now suppose x_1 and x_2 are independent and uniformly distributed on the interval $[\mu, 1]$. The density of a single random draw, x, is

$$f(x,\mu) = \frac{1}{1-\mu} 1_{\{\mu \leq x \leq 1\}}$$

where $1_{\{\mu \leq x \leq 1\}}$ is the indicator function on the set $[\mu, 1]$. This density function is log supermodular in (x,μ), because both $1_{\{\mu \leq x \leq 1\}}$ and $\frac{1}{1-\mu}$ are log supermodular in (x,μ), and the product of nonnegative log supmodular functions is log supermodular. Therefore $E(\mu|x_1,x_2)$ is weakly increasing in x_1 and x_2.

Let

$$\sigma(x_1,x_2) = \min(x_1,x_2)$$

The probability density of (x_1,x_2) can be written as follows:

$$f(x_1,\mu)f(x_2,\mu) = \frac{1}{1-\mu} 1_{\{\mu \leq x_1 \leq 1\}} \frac{1}{1-\mu} 1_{\{\mu \leq x_2 \leq 1\}} = \left(1_{\{0 \leq \min(x_1,x_2)\}}\right) \left(\frac{1}{(1-\mu)^2} 1_{\{\min(x_1,x_2) \geq \mu\}}\right)$$

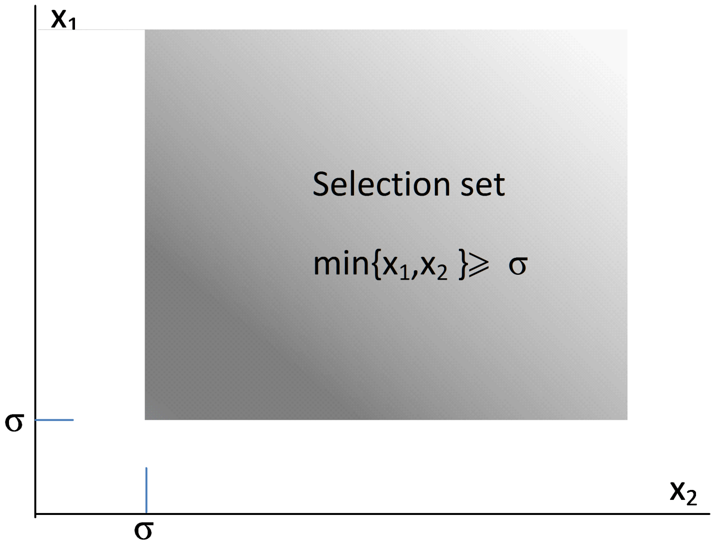


Figure 2.3: Selection for minimizing downside risk

Therefore, σ is sufficient for μ and there is an unbiased estimator of μ, β_2, that increases with σ. For this estimator,

$$E(\mu|x_1, x_2) = \beta_2(\min(x_1, x_2))$$

For some number α, an upper contour set of $E(\mu|x_1, x_2)$, and therefore the selection set Δ, takes the following form:

$$\Delta = \{(x_1, x_2)|\min(x_1, x_2) \geq \alpha\}$$

2.4 Double Rounds of elimination with memory

We now suppose that it is costly to collect information in each round, for example, because assistant professors must be paid. In the previous sections, we implicitly assumed that the selection will be made only after two draws. If drawing samples is not costly, this is optimal. However, when sampling is costly, money can be saved by discarding some unpromising projects or agents after the first round. That is, it is optimal to have double rounds of elimination. The potential penalty for saving money in this way is that the first round of elimination might exclude good projects that would be revealed as such if kept for the second round.

We assume that the cost of experimenting is the same in each round, namely, c. This is without loss of generality, assuming it is efficient to begin the experimentation process at all. Since the cost in the first round must be sunk in order to proceed, the (relevant) objective function depends only on the cost in the second round.

The selection process *has memory* if the selection criterion at round two can depend on the signal generated at round one. We view the selection problem as the choice of $\Delta_1 \subset \mathbf{R}$ and $\Delta_2 : \Delta_1 \to \mathcal{A}$, where $\mathcal{A}$ is the set of measurable subsets of $\mathbf{R}$ and where, for each $x_1 \in \Delta_1$, $\Delta_2(x_1) \in \mathcal{A}$ is understood as the selection set at the second round.

The objective is to maximize the expected μ among agents who survive both rounds, minus the cost that must be paid in the second round for survivors of the first round. The number of survivors at the end of the second round is constrained to be Λ. We write the objective function as

$$\begin{aligned} V(\Delta_1, \Delta_2; c) &\equiv \int \int_{\Delta_1} \int_{\Delta_2(x_1)} \mu p(x_1, x_2, \mu)\, dx_2 dx_1 d\mu \\ &\quad - c \int \int_{\Delta_1} \int \mu p(x_1, x_2, \mu)\, dx_2 dx_1 d\mu \end{aligned}$$

We write the number of survivors of both rounds as

$$S(\Delta_1, \Delta_2) =: \int \int_{\Delta_1} \int_{\Delta_2(x_1)} p(x_1, x_2, \mu)\, dx_2 dx_1 d\mu$$

Then the objective is

$$\text{maximize } V(\Delta_1, \Delta_2; c) \text{ subject to } S(\Delta_1, \Delta_2) = \Lambda \tag{2.8}$$

Theorem 2.4.1 (Double elimination with memory: the optimal policy) *Let Δ_1, Δ_2 be the selection sets that solve (2.8). Then there exists a number λ such that*

$$\begin{aligned} E(\mu | x_1, \Delta_2(x_1)) - \frac{c}{p(\Delta_2(x_1) | x_1)} - \lambda &\geq 0 \text{ for a.e. } x_1 \in \Delta_1 \\ E(\mu | x_1, \Delta_2(x_1)) - \frac{c}{p(\Delta_2(x_1) | x_1)} - \lambda &\leq 0 \text{ for a.e. } x_1 \notin \Delta_1 \end{aligned} \tag{2.9}$$

$$\text{for a.e. } x_1 \in \Delta_1, \begin{cases} E(\mu | x_1, x_2) - \lambda \geq 0 \text{ for a.e. } x_2 \in \Delta_2(x_1) \\ \\ E(\mu | x_1, x_2) - \lambda \leq 0 \text{ for a.e. } x_2 \notin \Delta_2(x_1) \end{cases} \tag{2.10}$$

Proof: In order to characterize the solution, it is convenient to reformulate the objective as a Lagrange function where the choice variables are indicator values $I(x_1) \in \{0, 1\}$ for each $x_1 \in \mathbf{R}$ and $J(x_1, x_2) \in \{0, 1\}$ for each $(x_1,x_2) \in \mathbf{R} \times \mathbf{R}$

such that $I(x_1) = 1$ There is a one-to-one relationship between the indicator functions I, J and the selection sets Δ_1, Δ_2, given by

$$\Delta_1 = \{x_1 \in \mathbf{R} : I(x_1) = 1\}$$

$$\Delta_2(x_1) = \{x_2 \in \mathbf{R} : J(x_1, x_2) = 1\} \text{ for each } x_1 \in \Delta_1$$

We will sometimes refer to the optimum as the optimal indicator functions I, J and sometimes as the optimal selection sets Δ_1, Δ_2.

The Lagrange function to be maximized is

$$\begin{aligned} &\mathcal{L}(I, J) = \int\int\int I(x_1) J(x_1, x_2) \quad \mu \quad p(x_1, x_2, \mu) \, dx_2 dx_1 d\mu \\ &-c \int\int\int I(x_1) \quad p(x_1, x_2, \mu) \, dx_2 dx_1 d\mu \\ &-\lambda \left[\int\int\int I(x_1) J(x_1, x_2) \quad p(x_1, x_2, \mu) \, dx_2 dx_1 d\mu - \Lambda \right] \end{aligned}$$

The first-order conditions are $S(\Delta_1, \Delta_2) = \Lambda$ and for each (x_1, x_2),

$$\begin{cases} \frac{\partial}{\partial I(x_1)} \mathcal{L}(I, J) \geq 0 & \text{if} \quad I(x_1) = 1 \\ \\ \frac{\partial}{\partial I(x_1)} \mathcal{L}(I, J) \leq 0 & \text{if} \quad I(x_1) = 0 \end{cases} \tag{2.11}$$

$$\begin{cases} \frac{\partial}{\partial J(x_1,x_2)} \mathcal{L}(I, J) \geq 0 & \text{if} \quad J(x_1, x_2) = 1 \\ \\ \frac{\partial}{\partial J(x_1,x_2)} \mathcal{L}(I, J) \leq 0 & \text{if} \quad J(x_1, x_2) = 0 \end{cases} \tag{2.12}$$

To interpret these conditions, we write out the values of the partial derivatives.

$$\begin{aligned} \frac{\partial}{\partial I(x_1)} \mathcal{L}(I, J) = \int\int J(x_1, x_2) \mu p(x_1, x_2, \mu) \, dx_2 d\mu - c \int\int p(x_1, x_2, \mu) \, dx_2 d\mu \\ -\lambda \int\int J(x_1, x_2) \, p(x_1, x_2, \mu) \, dx_2 d\mu \end{aligned}$$

Given a pair $(x_1, \Delta) \in \mathbf{R} \times \mathcal{A}$, we use the notation

$$E(\mu | x_1, \Delta) =: \int_\Delta \frac{p(x_2 | x_1)}{p(\Delta | x_1)} E(\mu | x_1, x_2) \, dx_2 \tag{2.13}$$

Then

$$\begin{aligned} \frac{\partial}{\partial I(x_1)} \mathcal{L}(I, J) &= \int J(x_1, x_2) \, p(x_1, x_2) \int \mu \frac{p(x_1, x_2, \mu)}{p(x_1, x_2)} d\mu dx_2 \\ &\quad -cp(x_1) - \lambda \int J(x_1, x_2) \, p(x_1, x_2) \, dx_2 \\ &= \int J(x_1, x_2) \, p(x_1, x_2) \, E(\mu | x_1, x_2) \, dx_2 - cp(x_1) - \lambda p(x_1, \Delta_2(x_1)) \end{aligned}$$

$$
\begin{aligned}
&\frac{\partial}{\partial I\left(x_1\right)}\mathcal{L}\left(I,J\right)\times\frac{1}{p\left(x_1,\Delta_2\left(x_1\right)\right)}\\
&=\int J\left(x_1,x_2\right)\frac{p\left(x_1,x_2\right)}{p\left(x_1,\Delta_2\left(x_1\right)\right)}E\left(\mu|x_1,x_2\right)dx_2-c\frac{p\left(x_1\right)}{p\left(x_1,\Delta_2\left(x_1\right)\right)}-\lambda\\
&=\int_{\Delta_2(x_1)}\frac{p\left(x_2|x_1\right)}{p\left(\Delta_2\left(x_1\right)|x_1\right)}E\left(\mu|x_1,x_2\right)dx_2-c\frac{1}{p\left(\Delta_2\left(x_1\right)|x_1\right)}-\lambda\\
&=E\left(\mu|x_1,\Delta_2\left(x_1\right)\right)-\frac{c}{p\left(\Delta_2\left(x_1\right)|x_1\right)}-\lambda \qquad (2.14)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial J\left(x_1,x_2\right)}\mathcal{L}\left(I,J\right) &= I\left(x_1\right)\left[p\left(x_1,x_2\right)\int_0^\infty\mu\frac{p\left(x_1,x_2,\mu\right)}{p\left(x_1,x_2\right)}d\mu-\lambda p\left(x_1,x_2\right)\right]\\
&= I\left(x_1\right)p\left(x_1,x_2\right)\times\left[E\left(\mu|x_1,x_2\right)-\lambda\right] \qquad (2.15)
\end{aligned}
$$

The conclusions in the theorem follow from (2.11) and (2.12), and (2.14) and (2.15). ■

The condition (2.9) for selection at the first round takes account of the cost that will be incurred in the second round. The value of keeping the agent after round one is diminished by the expected cost. The cost is wasted if the agent will be eliminated later. Averaged over the agents who survive both rounds, the per-agent cost of including the signal x_1 at the first round is $\frac{c}{p(\Delta_2(x_1)|x_1)}$.

At the second round of elimination, the decision maker can use both sample points (x_1, x_2) to select the ultimate survivors. Since both sample points contain information, the selection process should clearly use both. However, the selection is only among agents who survived round one – many sample points are "missing," and the ones that are missing were selected in a systematic way. This means that the conditional distribution of x_2, given the selection in round one, is different than the distribution of x_1 in round one, and different than the distribution of x_2 if all agents were in the sample. Given that there will be a "good luck bias" among the agents who survive round one, one might think that x_2 should be given some special weight in the evaluation at round two.

To put these questions more precisely,

- Should the two sample points (x_1, x_2) be treated symmetrically at the end of round two?
- For the case that there is a sufficient statistic for a single round of elimination, should the selection at the end of round two be based on the same statistic?

Perhaps surprisingly, the following corollary answers these questions affirmatively.

Corollary 2.4.2 (Double elimination with memory and a sufficient statistic) *Suppose σ is a sufficient statistic for μ. Let Δ_1, Δ_2 be the selection sets that solve (2.8). Then for a suitable constant λ,*

$$E\left(\mu|x_1, \Delta_2\left(x_1\right)\right) - \tfrac{c}{p(\Delta_2(x_1)|x_1)} - \lambda \geq 0 \textit{ for a.e. } x_1 \in \Delta_1$$

$$E\left(\mu|x_1, \Delta_2\left(x_1\right)\right) - \tfrac{c}{p(\Delta_2(x_1)|x_1)} - \lambda \leq 0 \textit{ for a.e. } x_1 \notin \Delta_1$$

$$\textit{for a.e. } x_1 \in \Delta_1, \begin{cases} \bar{E}\left(\mu|\sigma\left(x_1, x_2\right)\right) \geq \lambda \textit{ for a.e. } x_2 \in \Delta_2\left(x_1\right) \\ \\ \bar{E}\left(\mu|\sigma\left(x_1, x_2\right)\right) \leq \lambda \textit{ for a.e. } x_2 \notin \Delta_2\left(x_1\right) \end{cases}$$

Proof: The characterization of Δ_1 is the same as in Theorem 2.4.1, and the characterization of Δ_2 relies on the sufficient statistic instead of (x_1, x_2), using (2.6). ■

Thus, if there is a sufficient statistic for μ, the selection criterion at the second round depends only on this statistic. It is the same sufficient statistic as is used in a single round of elimination, even though the distributions are different. For example, if it is optimal to use only the sample mean for selection in a single round of elimination, then it is optimal to use the mean at round two – in particular, to weight the signals of the two periods equally – even when a prior selection has been made at round one, based only on x_1. No extra weight should be given to x_2 to compensate for the fact that x_1 has already been used at round one.

As with a single round of elimination, the characterization of the optimum is more useful if the density functions are log supermodular, and therefore satisfy the monotone likelihood ratio property. We now turn to this case.

2.4.1 Double round with memory: monotonicity and threshold values

With additional structure on the distributions, we can say more about how the signals (x_1, x_2) should be used in two rounds of elimination. In particular, the optimal selection sets (Δ_1, Δ_2) are threshold policies.

Theorem 2.4.3 shows that log supermodularity again leads to the conclusion that the optimal selection policy is a threshold policy in each round. Theorem 2.4.4 refines this result, showing that selection in the first round should use a threshold for the first signal, and if there is a sufficient statistic, should use a threshold for the sufficient statisic in the second round. Both these threshold results use log supermodularity.

Theorem 2.4.3 (Double elimination with memory: promotion using threshold values) *Suppose the distribution p is log supermodular. Given a cost c, let Δ_1, Δ_2 be the selection sets that solve (2.8). Then the optimal selection sets can be written with threshold values $k_1(c), \{k_2(x_1, c) : x_1 \geq k_1(c)\}$ such that*

$$x_1 \geq k_1(c) \text{ for a.e. } x_1 \in \Delta_1$$

$$x_1 \leq k_1(c) \text{ for a.e. } x_1 \notin \Delta_1$$

$$\text{for a.e. } x_1 \in \Delta_1, \begin{cases} x_2 \geq k_2(x_1, c) \text{ for a.e. } x_2 \in \Delta_2(x_1) \\ x_2 \leq k_2(x_1, c) \text{ for a.e. } x_2 \notin \Delta_2(x_1) \end{cases}$$

Proof: Taking Δ_2 first, log supermodularity implies that $E(\mu|x_1, x_2)$ is increasing in x_2. Therefore the conclusion follows from Theorem 2.4.1.

For Δ_1, referring to the proof of Theorem 2.4.1, rewrite the derivative (2.13) of the Lagrange function as

$$\begin{aligned} \frac{\partial}{\partial I(x_1)} \mathcal{L}(I, J) &= \int_{-\infty}^{\infty} J(x_1, x_2) p(x_1, x_2) [E(\mu|x_1, x_2) - \lambda] \, dx_2 - cp(x_1) \\ \frac{1}{p(x_1)} \frac{\partial}{\partial I(x_1)} \mathcal{L}(I, J) &= \int_{-\infty}^{\infty} J(x_1, x_2) p(x_2|x_1) [E(\mu|x_1, x_2) - \lambda] \, dx_2 - c \end{aligned}$$

By Theorem 2.4.1, $J(x_1, x_2) \geq 0 \iff [E(\mu|x_1, x_2) - \lambda] \geq 0$. Thus,

$$\frac{1}{p(x_1)} \frac{\partial}{\partial I(x_1)} \mathcal{L}(I, J) = \int_{-\infty}^{\infty} p(x_2|x_1) \max(E(\mu|x_1, x_2) - \lambda, 0) \, dx_2 - c$$

For clarity of the argument, define

$$\omega(x_1, x_2) =: \max(E(\mu|x_1, x_2) - \lambda, 0)$$

and write

$$\frac{\partial}{\partial I(x_1)} \mathcal{L}(I, J) = \int_{-\infty}^{\infty} p(x_2|x_1) \omega(x_1, x_2) \, dx_2 - c$$

To prove the result, we will show that $\frac{\partial}{\partial I(x_1)} \mathcal{L}(I, J) \leq \frac{\partial}{\partial I(\hat{x}_1)} \mathcal{L}(I, J)$ if $x_1 \leq \hat{x}_1$.

Due to log-supermodularity, $E(\mu|x_1, x_2)$ is increasing with both x_1 and x_2, and therefore ω is increasing in x_1 and x_2. To complete the proof, it is enough to that, if f is log supermodular and $x_1 \leq \hat{x}_1$, then

$$\int p(x_2|x_1) \max(E(\mu|x_1, x_2) - \lambda, 0) dx_2 \leq \int p(x_2|\hat{x}_1) \max(E(\mu|\hat{x}_1, x_2) - \lambda, 0) dx_2$$

Let

$$\begin{aligned} u(x_2) &= \max(E(\mu|\hat{x}_1, x_2) - \lambda, 0) \\ v(x_2) &= \max(E(\mu|x_1, x_2) - \lambda, 0) \\ F(x_2) &= \int_{\infty}^{x_2} p(\tilde{x}_2|\hat{x}_1) d\tilde{x}_2 \\ G(x_2) &= \int_{\infty}^{x_2} p(\tilde{x}_2|x_1) d\tilde{x}_2 \end{aligned}$$

In this notation, we want to show that

$$\int v(x_2)\, dG(x_2) \leq \int u(x_2)\, dF(x_2)$$

$E(\mu|x_1, x_2)$ is weakly increasing in both arguments by log supermodularity of $p(x_1, x_2, \mu)$, and therefore v and u are weakly increasing. Because $\hat{x}_1 > x_1$, $v \leq u$, and because F first-order dominates the distribution function G (see the Remark) the result follows from Lemma 2.3.3. ■

When there is a sufficient statistic, the optimal threshold policy can be stated with reference to the sufficient statistic in round two, just as for a single round of elimination.

Theorem 2.4.4 (Double elimination with memory: sufficient statistic) *Suppose the distribution p is log supermodular, and that σ is a sufficient statistic for μ. Given a cost c, let Δ_1, Δ_2 be the selection sets that solve (2.8). Then the selection sets can be written with threshold values $k_1, \bar{\sigma}$ such that*

$$x_1 \geq k_1 \text{ for a.e. } x_1 \in \Delta_1$$

$$x_1 \leq k \text{ for a.e. } x_1 \notin \Delta_1$$

$$\text{for a.e. } x_1 \in \Delta_1, \begin{cases} \sigma(x_1, x_2) \geq \bar{\sigma} \text{ for a.e. } x_2 \in \Delta_2(x_1) \\ \\ \sigma(x_1, x_2) \leq \bar{\sigma} \text{ for a.e. } x_2 \notin \Delta_2(x_1) \end{cases}$$

Proof: The characterization of Δ_1 is the same as in Theorem 2.4.3. We must show that the characterization of Δ_2 in Theorem 2.4.1 is equivalent to the one in this theorem. The posterior distribution of μ, given (x_1, x_2), is

$$\begin{aligned} p(\mu|x_1, x_2) &= \frac{p(x_1, x_2, \mu)}{\int p(x_1, x_2, \mu)\, d\mu} = \frac{g(x_1, x_2)\, \theta(\sigma(x_1, x_2), \mu)}{\int g(x_1, x_2)\, \theta(\sigma(x_1, x_2), \mu)\, d\mu} \\ &= \frac{\theta(\sigma(x_1, x_2), \mu)}{\int \theta(\sigma(x_1, x_2), \mu)\, d\mu} \end{aligned}$$

Therefore

$$\begin{aligned} E(\mu|x_1,x_2) &= \int \mu \left[\frac{\theta(\sigma(x_1,x_2),\mu)}{\int \theta(\sigma(x_1,x_2),\mu)\, d\mu} \right] d\mu \\ &= \bar{E}(\mu|\sigma(x_1,x_2)) \end{aligned}$$

Using Lemma 2.3.4, $\bar{E}(\mu|\sigma)$ is increasing in σ, and σ is increasing in (x_1, x_2). Choose $\bar{\sigma}$ so that $\bar{E}(\mu|\bar{\sigma}) = \lambda$. Then the characterization of Δ_2 above is equivalent to the characterization of Δ_2 in Theorem 2.4.1. ■

2.4.2 Double round with memory: comparisons

As shown in Theorem 2.4.1, the expected ability of the marginal survivor in an optimal selection scheme is λ at the end of round two. This is also the opportunity cost of reducing the number of survivors; it is the shadow price on the constraint that a fraction Λ of projects must survive. We now ask how the selection scheme and its efficacy change when the cost of keeping candidates in the pool increases. Given that mistakes are made at round one – some of the high-μ agents or projects are eliminated due to the randomness in x_1 – it is not entirely obvious how the ability of marginal survivors relates to the average ability in the group that survives.

The costliness of collecting information creates two burdens. First is the direct burden of paying the cost of round-one survivors in round two. Second, by eliminating some of the agents or projects too early, the selection process is less effective. We show that, if the cost of keeping survivors after round one increases, fewer will be kept, and the average ability of ultimate survivors, after round two, becomes smaller.

Theorem 2.4.5 (Double eliminations with memory: Higher cost leads to more stringent screening at round one, less stringent screening at round two, and lower average ability of survivors at the end.) *For each cost c, let (Δ_1^c, Δ_2^c) be the optimal selection sets for the double-elimination problem (2.8). Let $\hat{c} > c$. Then*
(1) If $0 < p\left(\Delta_1^{\hat{c}}\right) < 1$ and $0 < p(\Delta_1^c) < 1$, then $p\left(\Delta_1^{\hat{c}}\right) \leq p(\Delta_1^c)$.
(2) The expected ability of survivors in the selection scheme (Δ_1^c, Δ_2^c) is larger (no smaller) than in the selection scheme $\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}\right)$.
(3) If p is strictly log supermodular, the selection sets can be written with threshold values as in Theorem 2.4.3 where

$$\begin{aligned} k_1(\hat{c}) &\geq k_1(c) \\ k_2(x_1,\hat{c}) &\leq k_2(x_1,c). \end{aligned}$$

Proof: (1) Using Theorem 2.4.1, write the objective function as V, defined as

$$\begin{aligned} V\left(\Delta_1, \Delta_2; c\right) &= T\left(\Delta_1, \Delta_2\right) - cp\left(\Delta_1\right) \\ \text{where } T\left(\Delta_1, \Delta_2\right) &= \int \int_{\Delta_1} \int_{\Delta_2(x_1)} \mu p\left(x_1, x_2, \mu\right) dx_2 dx_1 d\mu \\ cp\left(\Delta_1\right) &= c \int \int_{\Delta_1} \int p\left(x_1, x_2, \mu\right) dx_2 dx_1 d\mu \end{aligned}$$

Because

$$\begin{aligned} V\left(\Delta_1^c, \Delta_2^c; c\right) &= T\left(\Delta_1^c, \Delta_2^c\right) - cp\left(\Delta_1^c\right) \geq T\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}\right) - cp\left(\Delta_1^{\hat{c}}\right) \\ V\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}; \hat{c}\right) &= T\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}\right) - \hat{c}p\left(\Delta_1^{\hat{c}}\right) \geq T\left(\Delta_1^c, \Delta_2^c\right) - \hat{c}p\left(\Delta_1^c\right) \end{aligned}$$

it follows that

$$\begin{aligned} T\left(\Delta_1^c, \Delta_2^c\right) - T\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}\right) &\geq c\left[p\left(\Delta_1^c\right) - p\left(\Delta_1^{\hat{c}}\right)\right] \\ T\left(\Delta_1^c, \Delta_2^c\right) - T\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}\right) &\leq \hat{c}\left[p\left(\Delta_1^c\right) - p\left(\Delta_1^{\hat{c}}\right)\right] \end{aligned}$$

Subtracting,

$$0 \geq \left(c - \hat{c}\right)\left[p\left(\Delta_1^c\right) - p\left(\Delta_1^{\hat{c}}\right)\right]$$

Thus $\left(c - \hat{c}\right) < 0 \Longrightarrow p\left(\Delta_1^c\right) \geq p\left(\Delta_1^{\hat{c}}\right)$.

(2) The total ability of survivors is $T\left(\Delta_1^c, \Delta_2^c\right)$ for each c. If $\left(\Delta_1^c, \Delta_2^c\right)$ is optimal for c, and $\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}\right)$ is optimal for $\hat{c}$, the two selection schemes yield the same number of survivors. Because $T\left(\Delta_1^c, \Delta_2^c\right) - cp\left(\Delta_1^c\right) \geq T\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}\right) - cp\left(\Delta_1^{\hat{c}}\right)$ and $p\left(\Delta_1^c\right) - p\left(\Delta_1^{\hat{c}}\right) \geq 0$, it follows that

$$T\left(\Delta_1^c, \Delta_2^c\right) - T\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}\right) \geq c\left[p\left(\Delta_1^c\right) - p\left(\Delta_1^{\hat{c}}\right)\right] \geq 0. \tag{2.16}$$

Dividing the left side by the probability of surviving both rounds, which is the same in both cases, the result follows.

(3) uses Theorem 2.4.3 and part (1) above.

$$\begin{aligned} p\left(\Delta_1^c\right) - p\left(\Delta_1^{\hat{c}}\right) &= \int \int_{k_1(c)} \int p\left(x_1, x_2, \mu\right) dx_2 dx_1 d\mu - \int \int_{k_1(\hat{c})} \int p\left(x_1, x_2, \mu\right) dx_2 dx_1 d\mu \\ &= \int \int_{k_1(c)}^{k_1(\hat{c})} \int p\left(x_1, x_2, \mu\right) dx_2 dx_1 d\mu \geq 0 \end{aligned}$$

The difference can only be positive if $k_1^{\hat{c}} \geq k_1\left(c\right)$, which also implies $k_2\left(x_1, \hat{c}\right) \leq k_2\left(x_1, c\right)$, because otherwise there would be different numbers of survivors in the two selection schemes. ■

Because a single round of elimination is the extreme case where $c = 0$ and everyone is promoted or rejected after one round, we have the following corollary:

Corollary 2.4.6 (When cost is zero, a single round of elimination is optimal) *Conditional on the same number of survivors, the expected ability of survivors after the optimal single round of elimination is larger than the expected ability of survivors in any double round of elimination in which some agents or projects are eliminated at round one.*

2.5 Double Rounds of elimination without memory

When we say that the elimination scheme does not have memory, we mean that the selection set at round two cannot depend on x_1. Only the fact of survival is known from the first round. The selection problem can now be described as the choice of $\Delta_1 \subset \mathbf{R}$ and $\Delta_2 \subset \mathbf{R}$, where the selection at round two requires both $x_1 \in \Delta_1$and $x_2 \in \Delta_2$.

The objective is still to maximize the expected μ among agents who survive both rounds, minus the cost that must be paid in the second round for survivors of the first round. We write the objective function as

$$\begin{aligned} V(\Delta_1, \Delta_2; c) &= :\int\int_{\Delta_1}\int_{\Delta_2} \mu p(x_1, x_2, \mu)\, dx_2 dx_1 d\mu \\ &\quad -c\int\int_{\Delta_1}\int \mu p(x_1, x_2, \mu)\, dx_2 dx_1 d\mu \end{aligned}$$

We write the number of survivors of both rounds as

$$S(\Delta_1, \Delta_2) =: \int\int_{\Delta_1}\int_{\Delta_2} p(x_1, x_2, \mu)\, dx_2 dx_1 d\mu$$

Then the objective is

$$\text{maximize } V(\Delta_1, \Delta_2; c) \text{ subject to } S(\Delta_1, \Delta_2) = \Lambda \tag{2.17}$$

Theorem 2.5.1 (Double elimination without memory: the optimal policy) *Let Δ_1, Δ_2 be selection sets that solve (2.17). Then there exists a number λ such that*

$$\begin{aligned} E(\mu|x_1, \Delta_2) - \frac{c}{p(\Delta_2|x_1)} - \lambda &\geq 0 \text{ for a.e. } x_1 \in \Delta_1 \\ E(\mu|x_1, \Delta_2) - \frac{c}{p(\Delta_2|x_1)} - \lambda &\leq 0 \text{ for a.e. } x_1 \notin \Delta_1 \end{aligned} \tag{2.18}$$

$$\begin{aligned} E(\mu|\Delta_1, x_2) - \lambda &\geq 0 \text{ for a.e. } x_2 \in \Delta_2 \\ E(\mu|\Delta_1, x_2) - \lambda &\leq 0 \text{ for a.e. } x_2 \notin \Delta_2 \end{aligned} \tag{2.19}$$

Proof: In order to characterize the solution, it is again convenient to reformulate the objective as a Lagrange function where the choice variables are indicator values $I(x_1) \in \{0,1\}$ for each $x_1 \in \mathbf{R}$ and $J(x_2) \in \{0,1\}$ for each $(x_1,x_2) \in \mathbf{R} \times \mathbf{R}$ such that $I(x_1) = 1$. There is again a one-to-one relationship between the indicator functions I, J and the selection sets Δ_1, Δ_2, now given by

$$\Delta_1 = \{x_1 \in \mathbf{R} : I(x_1) = 1\}$$

$$\Delta_2 = \{x_2 \in \mathbf{R} : J(x_2) = 1\}$$

The Lagrange function to be maximized is

$$\begin{aligned}
&\mathcal{L}(I,J) = \int\int I(x_1) \int J(x_2) \quad \mu\, p(x_1,x_2,\mu)\, dx_2 dx_1 d\mu \\
&-c \int I(x_1) \int\int p(x_1,x_2,\mu)\, dx_2 dx_1 d\mu \\
&-\lambda \left[\int\int I(x_1) \int J(x_2)\, p(x_1,x_2,\mu)\ \ dx_2 dx_1 d\mu - \Lambda\right]
\end{aligned}$$

The first-order conditions are $S(\Delta_1,\Delta_2) = \Lambda$ and for each (x_1,x_2),

$$\begin{cases} \frac{\partial}{\partial I(x_1)}\mathcal{L}(I,J) \geq 0 & \text{if} \quad I(x_1) = 1 \\ \\ \frac{\partial}{\partial I(x_1)}\mathcal{L}(I,J) \leq 0 & \text{if} \quad I(x_1) = 0 \end{cases} \tag{2.20}$$

$$\begin{cases} \frac{\partial}{\partial J(x_2)}\mathcal{L}(I,J) \geq 0 & \text{if} \quad J(x_2) = 1 \\ \\ \frac{\partial}{\partial J(x_2)}\mathcal{L}(I,J) \leq 0 & \text{if} \quad J(x_2) = 0 \end{cases} \tag{2.21}$$

$$\begin{aligned}
\frac{\partial}{\partial I(x_1)}\mathcal{L}(I,J) = \quad & \int\int J(x_2)\,\mu \int_{-\infty}^{\infty} p(x_1,x_2,\mu)\, dx_2 d\mu - c \int\int_{-\infty}^{\infty} p(x_1,x_2,\mu)\, dx_2 d\mu \\
& -\lambda \int\int J(x_2)\, p(x_1,x_2,\mu)\, dx_2 d\mu
\end{aligned}$$

For each $\Delta_1 \subset \mathbf{R}, \Delta_2 \subset \mathbf{R}$, we use the notation

$$\begin{aligned}
E(\mu|x_1,x_2) &= \int \mu \frac{p(x_1,x_2,\mu)}{\int p(x_1,x_2,\mu)\, d\mu} d\mu \\
&= \int \mu p(\mu|x_1,x_2)\, d\mu \\
E(\mu|x_1,\Delta_2) &= \int_{\Delta_2}\int \mu \frac{p(x_1,x_2,\mu)}{\int_{\Delta_2}\int p(x_1,x_2,\mu)\, d\mu dx_2} d\mu dx_2
\end{aligned}$$

$$\begin{aligned}
&= \int_{\Delta_2} E\left(\mu|x_1,x_2\right)\frac{p\left(x_2|x_1\right)}{p\left(\Delta_2|x_1\right)}dx_2 \\
E\left(\mu|\Delta_1,x_2\right) &= \int_{\Delta_1}\int \frac{p\left(x_1,x_2,\mu\right)}{\int_{\Delta_1}\int p\left(x_1,x_2,\mu\right)d\mu dx_1}d\mu dx_1 \\
&= \int_{\Delta_1} E\left(\mu|x_1,x_2\right)\frac{p\left(x_1|x_2\right)}{p\left(\Delta_1|x_2\right)}dx_1
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial I\left(x_1\right)}\mathcal{L}\left(I,J\right) &= \int J\left(x_2\right)p\left(x_1,x_2\right)\int \mu\frac{p\left(x_1,x_2,\mu\right)}{p\left(x_1,x_2\right)}d\mu dx_2 \\
&\quad -cp\left(x_1\right)-\lambda\int J\left(x_2\right)p\left(x_1,x_2\right)dx_2 \\
&= \int J\left(x_2\right)p\left(x_1,x_2\right)E\left(\mu|x_1,x_2\right)dx_2 - cp\left(x_1\right)-\lambda p\left(x_1,\Delta_2\right)
\end{aligned}$$

$$\frac{\partial}{\partial I\left(x_1\right)}\mathcal{L}\left(I,J\right)\times\frac{1}{p\left(x_1\right)} = \int J\left(x_2\right)p\left(x_2|x_1\right)E\left(\mu|x_1,x_2\right)dx_2 - c - \lambda\int J\left(x_2\right)p\left(x_2|x_1\right)dx_2 \tag{2.22}$$

$$\begin{aligned}
&\frac{\partial}{\partial I\left(x_1\right)}\mathcal{L}\left(I,J\right)\times\frac{1}{p\left(x_1,\Delta_2\right)} \\
&= \int J\left(x_2\right)\frac{p\left(x_1,x_2\right)}{p\left(x_1,\Delta_2\right)}E\left(\mu|x_1,x_2\right)dx_2 - c\frac{p\left(x_1\right)}{p\left(x_1,\Delta_2\right)}-\lambda \\
&= \int_{\Delta_2}\frac{p\left(x_2|x_1\right)}{p\left(\Delta_2|x_1\right)}E\left(\mu|x_1,x_2\right)dx_2 - c\frac{1}{p\left(\Delta_2|x_1\right)}-\lambda \\
&= E\left(\mu|x_1,\Delta_2\right)-\frac{c}{p\left(\Delta_2|x_1\right)}-\lambda
\end{aligned}\tag{2.23}$$

$$\begin{aligned}
\mathcal{L}\left(I,J\right) = &\int\int I\left(x_1\right)\int J\left(x_2\right)\quad\mu\quad p\left(x_1,x_2,\mu\right)dx_2dx_1d\mu \\
&-c\int I\left(x_1\right)\int\quad f\left(x_1,\mu\right)h\left(\mu\right)dx_1d\mu \\
&-\lambda\int\int I\left(x_1\right)\int J\left(x_2\right)\quad p\left(x_1,x_2,\mu\right)dx_2dx_1d\mu
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial J\left(x_2\right)}\mathcal{L}\left(I,J\right) &= \int I\left(x_1\right)\int\quad\mu\quad p\left(x_1,x_2,\mu\right)d\mu dx_1 - \lambda\int I\left(x_1\right)\int p\left(x_1,x_2,\mu\right)d\mu dx_1 \\
&= p\left(\Delta_1,x_2\right)\left\{E\left(\mu|\Delta_1,x_2\right)-\lambda\right\}
\end{aligned}\tag{2.24}$$

The theorem then follows from (2.20) and (2.21), and (2.23) and (2.24). ■

As with memory, if the probability distributions satisfy a monotonicity property, then the optimal selection sets can be expressed using threshold values, and the threshold values depend on the cost.

Theorem 2.5.2 [Double elimination without memory: promotion using threshold values] *Suppose the distribution p is log supermodular. Given a cost c, let Δ_1, Δ_2 be the selection sets that solve (2.17). Then the selection sets can be written with threshold values $k_1(c), k_2(c)$ such that*

$$x_1 \geq k_1(c) \text{ for a.e. } x_1 \in \Delta_1$$

$$x_1 \leq k_1(c) \text{ for a.e. } x_1 \notin \Delta_1$$

$$x_2 \geq k_2(c) \text{ for a.e. } x_2 \in \Delta_2$$

$$x_2 \leq k_2(c) \text{ for a.e. } x_2 \notin \Delta_2$$

This is proved as in Theorem 2.4.3.

Finally, we can make a qualitative statement about the stringency of screening at round one. Even without log supermodularity, we can conclude that higher cost should lead to fewer survivors of round one, and that the more stringent policy reduces the average ability of ultimate survivors. With log supermodularity, the more stringent policy takes the form of a higher threshold value for survival at round one, and a corresponding lower threshold at round two, in order to ensure that there are enough ultimate survivors.

Theorem 2.5.3 [Double eliminations without memory: Higher cost leads to more stringent screening at round one, less stringent screening at round two, and survivors of lower average ability at the end.]
For each cost c, let Δ_1^c, Δ_2^c be the selection sets that solve (2.17). Let $\hat{c} > c$. Then
(1) If $0 < p\left(\Delta_1^{\hat{c}}\right), p\left(\Delta_1^c\right) < 1$, then $p\left(\Delta_1^c\right) \geq p\left(\Delta_1^{\hat{c}}\right)$.
(2) The expected ability of survivors in the selection scheme $\left(\Delta_1^c, \Delta_2^c\right)$ is larger than in the selection scheme $\left(\Delta_1^{\hat{c}}, \Delta_2^{\hat{c}}\right)$.
(3) If p is strictly log supermodular, the selection sets can be written with threshold values as in Theorem 2.5.2, where

$$k_1(\hat{c}) > k_2(c) \text{ and } k_2(\hat{c}) \leq k_2(c).$$

The proof of this theorem is the same as the proof of Theorem 2.4.5 except that we must use the definition of V in the problem (2.17) where the selection criterion in round two does not depend on x_1.

Finally, we ask whether selection standards should become tougher or more lenient over time. That is, should the standard in the second round be lower or higher than in the first round? We did not address this question for the selection scheme with memory, because the selection standards in the second round depend on the signal from the first round.

It is convenient to return to the underlying densities f and h, instead of the density function p. Let ϕ be the hazard rate of f:

$$\phi(k,\mu) \equiv \frac{f(k,\mu)}{[1-F(k,\mu)]}$$

The assumption under which we can rank the thresholds in the two rounds is that the cross partial is positive:

$$\frac{\partial^2}{\partial\mu\partial k}\log\phi(k,\mu) > 0 \tag{2.25}$$

Lemma 2.5.4 *Let $k_1, k_2 \in \mathbf{R}$. Suppose that (2.25) holds. Then*

$$E(\mu|k_1,(k_2,\infty)) \left\{\begin{array}{c} > \\ = \\ < \end{array}\right\} E(\mu|(k_1,\infty),k_2) \quad \begin{array}{l} \textit{if } k_1 > k_2 \\ \textit{if } k_1 = k_2 \\ \textit{if } k_1 < k_2 \end{array}$$

Proof: Write

$$\begin{aligned} p((k_1,\infty),k_2) &= \int_{k_1}\int f(x_1,\mu)\,f(k_2,\mu)\,h(\mu)\,d\mu dx_1 \\ &= \int \phi(k_2,\mu)[1-F(k_1,\mu)][1-F(k_2,\mu)]\,h(\mu)\,d\mu \\ p(k_1,(k_2,\infty)) &= \int \phi(k_1,\mu)[1-F(k_1,\mu)][1-F(k_2,\mu)]\,h(\mu)\,d\mu \end{aligned}$$

$$\begin{aligned} E(\mu|(k_1,\infty),k_2) &= \int \mu\frac{\phi(k_2,\mu)[1-F(k_1,\mu)][1-F(k_2,\mu)]\,h(\mu)}{p((k_1,\infty),k_2)}d\mu \\ &= \int \mu g^{k_2}(\mu)\,d\mu \end{aligned}$$

where g^{k_2} is a probability distribution defined for each μ by

$$g^{k_2}(\mu) = \frac{\phi(k_2,\mu)[1-F(k_1,\mu)][1-F(k_2,\mu)]\,h(\mu)}{p((k_1,\infty),k_2)}$$

Similarly,

$$\begin{aligned} E\left(\mu | k_1, (k_2, \infty)\right) &= \int \mu \frac{\phi\left(k_1, \mu\right)\left[1 - F\left(k_1, \mu\right)\right]\left[1 - F\left(k_2, \mu\right)\right] h\left(\mu\right)}{p\left(k_1, (k_2, \infty)\right)} d\mu \\ &= \int \mu g^{k_1}\left(\mu\right) d\mu \end{aligned}$$

where g^{k_1} is a probability distribution defined for each μ by

$$g^{k_1}\left(\mu\right) = \frac{\phi\left(k_1, \mu\right)\left[1 - F\left(k_1, \mu\right)\right]\left[1 - F\left(k_2, \mu\right)\right] h\left(\mu\right)}{p\left(k_1, (k_2, \infty)\right)}$$

If $k_1 = k_2$, then $g^{k_1}\left(\mu\right) = g^{k_2}\left(\mu\right)$, hence $E\left(\mu | k_1, (k_2, \infty)\right) = E\left(\mu | \left(k_1, \infty\right), k_2\right)$.

If the cross partial (2.25) is positive, the following shows that the ratio $g^{k_1}\left(\mu\right)/g^{k_2}\left(\mu\right)$ is increasing if $k_1 > k_2$, and therefore g^{k_1} stochastically dominates g^{k_2}. The reverse holds if $k_2 > k_1$.

$$\begin{aligned} \frac{\partial}{\partial\mu} \log \frac{g^{k_1}\left(\mu\right)}{g^{k_2}\left(\mu\right)} &= \frac{\partial}{\partial\mu} \log \frac{\phi\left(k_1, \mu\right)}{\phi\left(k_2, \mu\right)} = \frac{\partial}{\partial\mu} \log \phi\left(k_1, \mu\right) - \frac{\partial}{\partial\mu} \log \phi\left(k_2, \mu\right) \\ &= \int_{k_2}^{k_1} \frac{\partial^2}{\partial\mu\partial k} \log \phi\left(k, \mu\right) dk \qquad \blacksquare \end{aligned}$$

This lemma allows us to state the following theorem:

Theorem 2.5.5 (Without memory, screening should become less stringent) *For a given cost* c, *suppose the selection sets that solve (2.17) can be written with threshold values* $\left(k_1\left(c\right), k_2\left(c\right)\right)$ *as in Theorem 2.5.2. Suppose that the cross partial of the derivative of the logarithm of* ϕ *is positive. Then for each* c, $k_1\left(c\right) > k_2\left(c\right)$.

Proof: Theorem 2.5.1 implies that

$$E\left(\mu | k_1\left(c\right), \left(k_2\left(c\right), \infty\right)\right) - \frac{c}{p\left(\left(k_2\left(c\right), \infty\right) | k_1\left(c\right)\right)} - E\left(\mu | \left(k_1\left(c\right), \infty\right), k_2\left(c\right)\right) = 0$$

If $c > 0$, the result follows from Lemma 2.5.4 because $E\left(\mu | k_1\left(c\right), \left(k_2\left(c\right), \infty\right)\right) > E\left(\mu | \left(k_1\left(c\right), \infty\right), k_2\left(c\right)\right)$. $\blacksquare$

Chapter 3

Appendix

3.1 Some Auxiliary Lemmas

A few technical lemmata are presented in this section, which may be skipped on first reading. Proofs of these results are in the following sections.

The first lemma lists some properties of g.

Lemma 3.1.1 *Let* $g(x,y) = \frac{x+\sqrt{x^2+4xy}}{2}, x > 0, y \geq 0$. *Then:*

1. *(Definition)* $g^2(x,y) = x(g(x,y)+y)$.
2. *(Homogeneity)* $g(\lambda x, \lambda y) = \lambda g(x,y), \forall \lambda \geq 0, x \geq 0, y \geq 0$.
3. *(Monotonicity)* $g(x,y)$ *is strictly increasing in* x *and* y, *and* $g_y = \frac{x}{2g-x} \leq 1 \leq g_x = \frac{g+y}{2g-x}$.
4. *(Bounds)* $x \leq g(x,y) \leq x+y$.
5. *(Concavity) If* $x > 0, y > 0$, *then* $g_{xx} < 0, g_{yy} < 0$ *and* $g_{xy} > 0$.
6. *(Special Values)* $g(x,0) = x$ *and* $g(\frac{y}{1+n}, ny) = y, \forall n, y$.

The next three lemmata are used in the proof of Theorem 1.4.14.

Lemma 3.1.2 *If* $x \geq 0$, *then* $x/2 \leq g(x, g(x,t)+t) - g(x,t) \leq x, \quad \forall t \geq 0$. *All inequalities are strict if* $x > 0$.

Lemma 3.1.3 *If* $k, y \geq 0$, *then* $g(k/3, y + g(2k/3, y)) \leq g(2k/3, y)$, *with strict inequality if* $y > 0$ *and* $k > 0$.

Lemma 3.1.4 *Suppose* $0 < k \leq 1, y \geq 0$. *Define two functions* $\tilde{a}, a : [0,k] \to \mathbf{R}$ *as follows:*

$$\tilde{a}(e) = g(k-e, g(e,y)+y), \qquad a(e) = g(e,y),$$

Let

$$A = \tilde{a} + a, \qquad M = \tilde{a} - \frac{1}{2}\tilde{a}^2 + a - \frac{1}{2}a^2$$

Then:

1. $A(k) = A(0) < A(e), \forall e \in (0,k)$, *and* $M(k) = M(0) < M(e), \forall e \in (0,k)$.

2. A *is strictly concave in* e *and* $A'(\frac{k}{2}) > 0$. *Also,* $A'(\frac{2k}{3}) < 0$ *if* $y > 0$, $A'(\frac{2k}{3}) = 0$ *if* $y = 0$*; Therefore* $A'(e) > 0$ *if* $e \in [0, \frac{k}{2}]$, $A'(e) < 0$ *if* $e \in (\frac{2k}{3}, k]$*;*

3. $M'(e) > 0$ *on* $[0, \frac{k}{2}]$, *while* $M'(e) < 0$ *on* $(\frac{2k}{3}, k]$.

4. *If* $y = 0$, *then* $A'(\frac{2k}{3}) = M'(\frac{2k}{3}) = 0$, *and* $\frac{2k}{3}$ *is the* unique *maximizer of* A *and* M. *If* $y > 0$, *then* $A'(\frac{2k}{3}) < 0$, *and* $M'(\frac{2k}{3}) < 0$.

Lemma 3.1.5 *Suppose* $f : [0,\infty) \to \mathbf{R}$ *is continuous and differentiable, and* $f(0) = 0$. *If* $f'(x) \leq 0$ *whenever* $f(x) \geq 0$, *then* $f(x) \leq 0, \forall x \geq 0$.

Proof of Lemma 3.1.1

Proof Most of the calculations are straightforward.

1. If $x(1+\frac{y}{g}) = g$, then $g^2 = x(g+y)$, or $g^2 - xg - xy = 0$, hence $g = \frac{x+\sqrt{x^2+4xy}}{2}$ (drop the negative solution as $g > 0$).

2. Trivial.

3. If $x > 0, y > 0$, then g is differentiable in (x,y). Differentiating the equation $g^2 = x(g+y)$ with respect to x, we get $2gg_x = g + y + xg_x$, hence $g_x = \frac{g+y}{2g-x} > 0$. Similarly, we have $2gg_y = x(g_y+1)$, hence $g_y = \frac{x}{2g-x} > 0$. Note $x + y \geq g$ (part 4 below), so $g + y \geq 2g - x$, hence $g_x = \frac{g+y}{2g-x} \geq 1$. Similarly $g_y = \frac{x}{2g-x} \leq 1$ since $g \geq x$.

4. This follows from $x = \frac{x+\sqrt{x^2+0}}{2} \leq g(x,y) = \frac{x+\sqrt{x^2+4xy}}{2} \leq \frac{x+\sqrt{x^2+4xy+4y^2}}{2} = x + y$.

5. $g(x,y) = x\zeta(y/x)$, where $\zeta(z) = \frac{1+\sqrt{1+4z}}{2}$. It is easy to see that $\zeta''(z) < 0$, hence ζ is strictly concave. So, $g_{yy} = \frac{1}{x}\zeta''(y/x) < 0$ and $g_{xy} = -\frac{y}{x^2}\zeta''(y/x) > 0$, and $g_{xx} = \frac{y^2}{x^3}\zeta''(y/x) < 0$. Also $g_{xx}g_{yy} - g_{xy}^2 = 0$, therefore g is concave in (x,y).

6. Obviously $g(x,0) = x$. Also, $g(1, n(n+1)) = \frac{1+\sqrt{1+4n(n+1)}}{2} = \frac{1+2n+1}{2} = n+1$. By homogeneity of g, $g(\frac{y}{1+n}, ny) = \frac{y}{1+n}g(1, n(n+1)) = y$. ■

Proof of Lemma 3.1.2

Proof If $x = 0$, all terms vanish for any $t \geq 0$, hence the result holds in this case.

If $x > 0$, by the Mean Value Theorem (MVT),

$$g(x, g(x,t)+t) - g(x,t) = (g(x,t)+t-t)\, g_y(x,\zeta)$$

for $\zeta \in (t, t+g(x,t))$. Note g is concave in y, hence $g_y(x,\zeta) < g_y(x,t) = \frac{x}{2g(x,t)-x}$ (Part 3, Lemma A.1). Therefore

$$g(x, g(x,t)+t) - g(x,t) < g(x,t)g_y(x,t) = g(x,t)\frac{x}{2g(x,t)-x} = x\frac{g(x,t)}{2g(x,t)-x}.$$

Note $g(x,t) \geq x$, hence $g(x,t) \leq 2g(x,t) - x$, or equivalently $\frac{g(x,t)}{2g(x,t)-x} \leq 1$. So,

$$g(x, g(x,t)+t) - g(x,t) < x\frac{g(x,t)}{2g(x,t)-x} \leq x \cdot 1 = x$$

For the other direction, note $g_y(x,\zeta) > g_y(x, g(x,t)+t) = \frac{x}{2g(x,g(x,t)+t)-x}$, hence

$$g(x, g(x,t)+t) - g(x,t) > \frac{xg(x,t)}{2g(x,t+g(x,t))-x}$$

Simplifying this inequality, we have [1]

$$g(x, g(x,t)+t) - g(x,t) > \frac{x}{2} > \frac{xg(x,t)}{2g(x,t+g(x,t))-x}.$$

Combining both directions, we get the following chain of inequalities when $x > 0$:

$$\frac{xg(x,t)}{2g(x,t+g(x,t))-x} < x/2 < g(x,g(x,t)+t) - g(x,t) < x\frac{g(x,t)}{2g(x,t)-x} \leq x, \forall t \geq 0 \tag{3.1}$$

Hence Lemma 3.1.2 is proved. ■

Proof of Lemma 3.1.3

Proof If $k = 0$, then both sides equal zero for any y, hence the result holds.

If $k > 0$, by homogeneity of g, we have

$$g(\frac{k}{3}, g(\frac{2k}{3}, y) + y) - g(\frac{2k}{3}, y) = \frac{k}{3}\left(g(1, g(2, \frac{2y}{k}) + \frac{2y}{k}) - g(2, \frac{2y}{k})\right)$$

[1] $(\tilde{g} - g)(2\tilde{g} - x) > xg$, hence $2\tilde{g}^2 - 2g\tilde{g} - x\tilde{g} > 0$, or $\tilde{g}(2\tilde{g} - 2g - x) > 0$. Thus $2\tilde{g} - 2g - x > 0$ or $\tilde{g} - g > x/2$. Here $\tilde{g} = g(x, g(x,t)+t)$.

So, it is sufficient to show the case for $k = 3$. To this end, it is equivalent to show

$$z(y) := g(1, g(2, y) + y) - g(2, y) < 0, \forall y > 0.$$

Note $z'(y) = g_y(1, g(2, y) + y)(1 + g_y(2, y)) - g_y(2, y)$. Substituting for the partial derivative of g and simplifying, we have

$$z'(y) < 0 \Longleftrightarrow 2g(1, g(2, y) + y) - 1 - g(2, y) > 0.$$

Note $2g(1, g(2, y) + y) = g(2, 2(g(2, y) + y)) > g(2, g(2, y) + y) > 1 + g(2, y)$, $\forall y > 0$, while the last inequality follows from Lemma 3.1.2. So, $z'(y) < 0, \forall y > 0$, but $z(0) = g(1, 2) - g(2, 0) = 2 - 2 = 0$. Therefore, $z(y) < 0, \forall y > 0$. That finishes the proof of Lemma 3.1.3. ■

Proof of Lemma 3.1.4

Proof The proof is given in four parts.

Part 1:

Since $\tilde{a}(0) = a(k) = g(k, y), \tilde{a}(e) = a(k) = 0$. Therefore $A(k) = A(0)$, and $M(0) = M(k)$.

Suppose $0 < e < k$. Then [2]

$$\begin{aligned} A(e) &= g(k-e, g(e, y) + y) + g(e, y) > g(k-e, y) + g(e, y) \\ &\geq \frac{k-e}{k} g(k, y) + \frac{e}{k} g(k, y) = g(k, y) = A(0). \end{aligned}$$

For M, first simplify the expression using $g^2(x, y) = x(g(x, y) + y)$:

$$\begin{aligned} M(e) &= \tilde{a}(e) - \frac{1}{2}(k-e)(\tilde{a}(e) + a(e) + y) + a(e) - \frac{1}{2}e(a(e) + y) \\ &= (1 - \frac{k}{2})(\tilde{a}(e) + a(e)) + \frac{1}{2}e\tilde{a}(e) - \frac{1}{2}ky \\ &= (1 - \frac{k}{2})(\tilde{a}(e) + a(e)) + \frac{1}{2}\sqrt{e^2(k-e)(\tilde{a}(e) + a(e) + y)} - \frac{1}{2}ky \\ &= (1 - \frac{k}{2})A(e) + \frac{1}{2}\sqrt{e^2(k-e)(A(e) + y)} - \frac{1}{2}ky \qquad (3.2) \end{aligned}$$

If $0 < e < k$, then $A(e) > A(0) > 0$. Also $\sqrt{e^2(k-e)(A(e) + y)}$ is zero if $e = 0$, or $e = k$ and is positive $\forall e \in (0, k)$. Hence $M(e) > M(0)$.

Part 2:

For brevity, I use $\tilde{g} := g(k-e, g(e, y) + y), g := g(e, y)$. Then,

$$\begin{aligned} A'(e) &= -g_x(k-e, g(e, y) + y) + g_y(k-e, g(e, y) + y)g_x(e, y) + g_x(e, y) = -\tilde{g}_x + (1 + \tilde{g}_y)g_x \\ A''(e) &= \tilde{g}_{xx} - \tilde{g}_{xy}g_x + (-\tilde{g}_{yx} + \tilde{g}_{yy}g_x)g_x + (1 + \tilde{g}_y)g_{xx} \end{aligned}$$

[2] Note $g(e, y)$ is concave in e with $g(0, y) = 0$, hence $\frac{g(e,y)}{e}$ is weakly decreasing in $e \in [0, k]$.

Obviously $A''(e) < 0$, because $\tilde{g}_{xx} < 0, \tilde{g}_{xy} > 0, \tilde{g}_{yy} < 0, g_x > 0, \tilde{g}_y \geq 0$, so A is strictly concave. Notice that $A(0) = A(k)$, hence A has a unique interior maximizer on $[0, k]$, which is given by the solution to $A'(e) = 0$. Note that

$$A'(e) > 0 \Longleftrightarrow g_x > \frac{\tilde{g}_x}{1 + \tilde{g}_y}$$

From Lemma 3.1.1, $g_x = \frac{g+y}{2g-x}, \frac{\tilde{g}_x}{1+\tilde{g}_y} = \frac{\tilde{g}+g+y}{2\tilde{g}}$. Therefore,

$$\begin{aligned}
& g_x = \frac{g+y}{2g-e} > \frac{\tilde{g}_x}{1+\tilde{g}_y} = \frac{\tilde{g}+g+y}{2\tilde{g}} \\
& \Longleftrightarrow (g+y)2\tilde{g} > (2g-e)(\tilde{g}+g+y) = \tilde{g}(2g-e) + (g+y)(2g-e) \\
& \Longleftrightarrow \tilde{g}(2y+e) > (g+y)(2g-e) = 2g^2 + g(2y-e) - ey \\
& \qquad\quad = 2e(g+y) + g(2y-e) - ey = g(2y+e) + ey \\
& \Longleftrightarrow (\tilde{g}-g)(2y+e) > ey.
\end{aligned}$$

In the end, we have

$$A'(e) > (<)0 \Longleftrightarrow \tilde{g} - g > (<)\frac{ey}{2y+e} \tag{3.3}$$

When $e = \frac{k}{2}$, $\tilde{g} - g = g(k/2, g(k/2, y) + y) - g(k/2, y) > k/4$ by Lemma 3.1.2. Meanwhile $\frac{ey}{2y+e} = \frac{y}{2y+e}e < \frac{1}{2}e = k/4$. Therefore $A'(\frac{k}{2}) > 0$. By concavity, $A'(e) > 0$ if $e \in [0, \frac{k}{2}]$.

When $e = \frac{2k}{3}$, there are two cases. If $y = 0$, then $\tilde{g} = g(k/3, \frac{2k}{3}) = \frac{2k}{3} = g$. Therefore $A'(\frac{2k}{3}) = 0$ by equation (3.3). If $y > 0$, $A'(\frac{2k}{3}) < 0$ follows from $\tilde{g} - g = g(\frac{k}{3}, g(\frac{2k}{3}, y) + y) - g(\frac{2k}{3}, y) < 0$, by Lemma 3.1.3.

Therefore $A'(\frac{2k}{3}) < 0$, if $y > 0$; $A'(\frac{2k}{3}) = 0$, if $y = 0$.

Part 3:

By equation (3.2),

$$M'(e) = (1 - \frac{k}{2})A'(e) + \frac{1}{4}\{e^2(k-e)(A(e)+y)\}^{-1/2}\{e^2(k-e)(A(e)+y)\}'$$

where $\{e^2(k-e)(A(e)+y)\}' = e(2k-3e)(A(e)+y) + e^2(k-e)A'(e)$.

If $e \in [0, \frac{k}{2}]$, then $A'(e) > 0$ (by part 2) and $2k - 3e > 0$, so $\{e^2(k-e)(A(e)+y)\}' > 0$, hence $M'(e) > 0$ on $[0, \frac{k}{2}]$. Similarly, if $e \in (\frac{2k}{3}, k]$, then $A'(e) < 0$ (by part 2) and $2k - 3e < 0$, so $\{e^2(k-e)(A(e)+y)\}' < 0$. Thus $M'(e) < 0$ on $(\frac{2k}{3}, k]$.

Part 4:

If $y = 0$, at $e = \frac{2k}{3}$, $A'(\frac{2k}{3}) = 0$, so $\{e^2(k-e)(A(e)+y)\}'|_{e=\frac{2k}{3}} = 0$ (both $A'(e)$ and $2k - 3e$ vanish at this point), therefore $M'(\frac{2k}{3}) = 0$. Also, in this case, if $e < \frac{2k}{3}$, $A'(e) > 0$, hence $M'(e) > 0$. If $e > \frac{2k}{3}$, then $A'(e) < 0$, hence $M'(e) < 0$. Thus $\frac{2k}{3}$ is the unique maximizer of both M and A on $[0, k]$.

If $y > 0$, then at $e = \frac{2k}{3}$, $A'(\frac{2k}{3}) < 0$. Therefore $\{e^2(k-e)(A(e)+y)\}'|_{e=\frac{2k}{3}} < 0$ (note $2k - 3e$ vanishes at this point), hence $M'(\frac{2k}{3}) < 0$. Moreover, M is decreasing on $(\frac{2k}{3}, k]$ by part 3, so the maximizer of M is less than $\frac{2k}{3}$. ■

Proof of Lemma 3.1.5

Proof For $\epsilon > 0$, let $g(x) = f(x) - \epsilon(1+x)$. If $g(x) \geq 0$, then $f(x) \geq 0$, and $f'(x) \leq 0$. Hence $g'(x) = f'(x) - \epsilon < 0$. We claim that $g(x) \leq 0$ for all $x \geq 0$. Suppose $g(\bar{x}) > 0$. Then let $\hat{x} = \inf\{x \geq 0 | g(x) \geq 0\}$. Note $g(0) < 0$, hence $\hat{x} \neq 0$. Moreover $g(\hat{x}) = 0$ and $g(x) < 0$ for $\forall x < \hat{x}$. Therefore $g'(\hat{x}) \geq 0$. Also, $g(\hat{x}) = 0$, so $g'(\hat{x}) < 0$, hence we get a contradiction. Therefore $g(x) \leq 0, \forall x$. This implies $f(x) \leq \epsilon(1+x)$. This holds for any positive ϵ. Taking the limit as ϵ goes to zero, we have $f(x) \leq 0, \forall x$. ■

3.2 Omitted Proofs

All omitted proofs are given in this section.

Proof of Theorem 1.4.1

Before the proof, we need an auxiliary lemma.

Lemma 3.2.1 *For a simple hierarchy $\mathcal{H}$, the following are true:*

A. $\{DF^j : j \in N^k\}$ is a partition of N^{k+1}, for $k = 1, 2, \cdots, h-1$; that is, $\cup_{j \in N^k} DF^j = N^{1+k}$ and for $j \neq i, DF^j \cap DF^i = \emptyset$.

B. If $i, j \in N^k$ and $i \neq j$, then $F^i \cap F^j = \emptyset$.

C. For any two members i, j, there is at most one path from i to j.

Proof For A, the union of DF^j is N^{k+1} by part (b) of Definition 3. The disjointness of these sets follows from simplicity of the hierarchy. B and C follow from A using induction. ■

Proof of Theorem 1.4.1 If $\mathcal{H}$ is simple, then every worker except the leaders has a unique predecessor, hence a unique source of information. Also, F^i identifies the set of players whose beliefs can be influenced by i's effort. Lemma 3.2.1 shows that if $i, j \in N^k$, then $F^i \cap F^j = \emptyset$, i.e, i and j have no common followers, which makes the equilibrium characterization much easier.

We are interested in separating equilibrium, in which any player's effort reveals his belief about the state. For each player $i \in \mathcal{N}$, let $\tilde{x}_i : \Theta \to [0, +\infty)$ denote player i's optimal effort given his belief about the state. The equilibrium condition is that for any i,

$$\tilde{x}_i(\theta) \in \arg\max_{x_i \in \mathbf{R}^+} \; s_i\theta \left(x_i + \sum_{j \in F^i} \tilde{x}_j(\tilde{x}_i^{-1}(x_i)) \right) - c(x_i). \tag{3.4}$$

The first-order condition for equation 3.4 is

$$s_i\theta \left(1 + \sum_{j \in F^i} \frac{\tilde{x}_j'(\tilde{x}_i^{-1}(x_i))}{\tilde{x}_i'(\tilde{x}_i^{-1}(x_i))} \right) - c'(x_i) = 0, \text{ when } x_i = \tilde{x}_i(\theta).$$

Note $\tilde{x}_i^{-1}(x_i) = \theta$ if $x_i = \tilde{x}_i(\theta)$. Simplifying the above expression, we get:

$$s_i\theta \left(1 + \sum_{j \ \in \ F^i} \frac{\tilde{x}_j'(\theta)}{\tilde{x}_i'(\theta)} \right) = c'(\tilde{x}_i(\theta)).$$

This must hold for any $\theta \in \Theta$, which is just equation 1.2. ■

Remark 5 *In equation 3.4, we only consider the contributions of players in F^i and i, but ignore the contributions of others workers. The reason is other workers cannot be influenced by i's effort, hence their contributions only affect i's equilibrium payoff and do not affect i's incentive for signaling. By Lemma 3.2.1, we can isolate player i's problem from other players on the same level because they have disjoint sets of followers.*

Remark 6 *In general, we need to specify initial conditions to solve for ordinary differential equations. We do not need to do so here because $\tilde{x}_i(0) = 0, \forall i \in \mathcal{N}$ is implicitly implied by equation 1.2 by setting $\theta = 0$. If $\min \Theta = \underline{\theta} > 0$, then the initial condition for equation 1.2 (see Mailath, 1987) is fixed by requiring that the "worst" type, $\underline{\theta}$, get his maximal utility given that he is identified as the worst type; in other words, $\tilde{x}_i(\underline{\theta}) = c'^{-1}(s_i\underline{\theta})$. In general, no explicit solutions exist when $\underline{\theta} > 0$ even with quadratic disutility function.*

Proof of Lemma 1.4.8:

Part 1 Let ρ be the inverse map of σ. Then ρ is also a permutation of $\mathcal{N}$. It suffices to show:

$$(+) \qquad k_{\rho(i)}(\mathbf{s}, \mathcal{C}^{\sigma}_{(\mathcal{N})}) \geq k_{\rho(i)}(\mathbf{s}, \mathcal{H}), \forall i \in \mathcal{N}$$

We prove (+) by induction on i from bigger i to smaller i.
For $i = N$, we know that $\rho(N)$ is the worker on the last level, hence has no followers in the chain. So $\sigma(\rho(N)) = N \geq \sigma(i), \forall i$. Therefore $\rho(N)$ has no follower under $\mathcal{H}$. Thus, $k_{\rho(i)}(\mathbf{s}, \mathcal{C}^{\sigma}_{(\mathcal{N})}) = s_{\rho(N)} = k_{\rho(i)}(\mathbf{s}, \mathcal{H})$.

Suppose (+) holds for all i greater than or equal to K. If $i = K - 1$, then by Theorem 4.3

$$k_{\rho(K-1)}(\mathbf{s}, \mathcal{H}) = g\left(s_{\rho(K-1)}, \sum_{j \in F^{\rho(K-1)}} k_j(\mathbf{s}, \mathcal{H})\right)$$

By monotonicity of σ, we have

$$F^{\rho(K-1)} \subset \{j | \sigma(j) > \sigma(\rho(K-1) = K-1\} = \{j | \sigma(j) \geq K\} = \{\rho(l) | l \geq K\}.$$

Therefore,

$$\begin{aligned} k_{\rho(K-1)}(\mathbf{s}, \mathcal{H}) &= g\left(s_{\rho(K-1)}, \sum_{j \in F^{\rho(K-1)}} k_j(\mathbf{s}, \mathcal{H})\right) \\ &\leq g\left(s_{\rho(K-1)}, \sum_{l \geq K} k_{\rho(l)}(\mathbf{s}, \mathcal{H})\right) \\ &\leq g\left(s_{\rho(K-1)}, \sum_{l \geq K} k_{\rho(l)}(\mathbf{s}, \mathcal{C}^{\sigma}_{(\mathcal{N})})\right) \qquad \text{(by induction)} \\ &= k_{\rho(K-1)}(\mathbf{s}, \mathcal{C}^{\sigma}_{(\mathcal{N})}) \end{aligned}$$

Therefore, (+) holds for $K-1$. By induction, (+) holds for any $i = 1, \cdots, N$.

Part 2 We construct $\tilde{\mathbf{s}}$ step-by-step to satisfy the following conditions:

$$(++) \qquad\qquad k_{\rho(i)}(\tilde{\mathbf{s}}, \mathcal{C}^{\sigma}_{(\mathcal{N})}) = k_{\rho(i)}(\mathbf{s}, \mathcal{H}), \forall i \in \mathcal{N}$$

For $i = N$, let $\tilde{s}_{\rho(N)} = s_{\rho(N)}$. Note $\rho(N)$ has no followers under $\mathcal{H}$ or $\mathcal{C}^{\sigma}_{(\mathcal{N})}$, so in this case, $k_{\rho(N)}(\tilde{\mathbf{s}}, \mathcal{C}^{\sigma}_{(\mathcal{N})}) = \tilde{s}_{\rho(N)} = s_{\rho(N)} = k_{\rho(i)}(\mathbf{s}, \mathcal{H})$. So (++) holds for $i = N$.
Suppose we have constructed $\tilde{s}_i$ for all $i \geq K$. Define $\epsilon \geq 0$ such that:

$$g\left(s_{\rho(K-1)}, \sum_{j \in F^{\rho(K-1)}} k_j(\mathbf{s}, \mathcal{H})\right) = g\left(s_{\rho(K-1)} - \epsilon, \sum_{l \geq K} k_{\rho(l)}(\mathbf{s}, \mathcal{H})\right).$$

This ϵ always exists by continuity of g, because the right hand side is bigger than the left hand side if $\epsilon = 0$, and the right hand side is zero if $\epsilon = s_{\rho(K-1)}$. Let $\tilde{s}_{\rho(K-1)} = s_{\rho(K-1)} - \epsilon \leq s_{\rho(K-1)}$. Then

$$\begin{aligned}
k_{\rho(K-1)}(\mathbf{s}, \mathcal{H}) &= g(s_{\rho(K-1)}, \sum_{j \in F^{\rho(K-1)}} k_j(\mathbf{s}, \mathcal{H})) \\
&= g(s_{\rho(K-1)} - \epsilon, \sum_{l \geq K} k_{\rho(l)}(\mathbf{s}, \mathcal{H})) \\
&= g(\tilde{s}_{\rho(K-1)}, \sum_{l \geq K} k_{\rho(l)}(\tilde{\mathbf{s}}, \mathcal{C}^{\sigma}_{(\mathcal{N})})) \quad \text{by induction} \\
&= k_{\rho(K-1)}(\tilde{\mathbf{s}}, \mathcal{C}^{\sigma}_{(\mathcal{N})})
\end{aligned}$$

Therefore, (++) holds for $i = K - 1$. The results follow by induction. ■

Proof of Theorem 1.4.9

Proof Suppose $\mathbf{s} = \{s_i, i \in \mathcal{N}\}$ is optimal for $\phi(t_2, \mathcal{H})$ (an optimum always exists by continuity of $w(\mathbf{s}, \mathcal{H})$ and compactness of Δ^N). Note $\sum_{i \in \mathcal{N}} s_i = t_1 < t_2$ and let $\Delta = t_2 - t_1 > 0$. Choose one terminal worker, say b. Suppose his share is $s_b \in [0, t_1]$. Obviously, b has no followers. Let P^b be the set of b's predecessors. Also, we have to remove the workers with zero shares, so define $M = \{j \in P^b | s_j > 0\}$. It is easy to see that $k_j(\mathbf{s}, \mathcal{H})$ is strictly increasing in s_b if $j \in M$.
Let $\hat{e} = \Delta + \sum_{j \in M} s_j > 0$. We claim that:

Claim 3.2.2 *There exist functions $\{f_j, j \in M\}$ defined on $e \in [0, \hat{e}]$ such that:*

(1) $\forall j \in M$, f_j is continuous and nonnegative in e with $f_j(0) = 0$.

(2) $s_j - f_j(e) > 0$.

(3) $k_j(\tilde{\mathbf{s}}(e)) = k_j(\mathbf{s}), j \neq b$, and $k_b(\tilde{\mathbf{s}}(e)) = s_b + e$, where $\tilde{\mathbf{s}}(e)$ is the shares derived from $\mathbf{s}$ by adjusting $s_b \to s_b + e$ and $s_j \to s_j - f_j(e), j \in M$, and keeping all other workers' shares fixed.

If this claim is true, then define $\eta(e) = \Delta - e + \sum_{j \in M} f_j(e), e \in [0, \hat{e}]$. Notice that the summation of shares for $\tilde{\mathbf{s}}(e)$ is

$$|\tilde{\mathbf{s}}(e)| = \sum_{i \in \mathcal{N}} s_j + e - \sum_{j \in M} f_j(e) = t_1 + e - \sum_{j \in M} f_j(e) = t_1 + \Delta - \eta(e) = t_2 - \eta(e),$$

which varies with e. Also, $\eta(0) = \Delta - 0 + 0 > 0$, and

$$\eta(e) = \Delta - e + \sum_{j \in M} f_j(e) \leq \Delta - e + \sum_{j \in M} s_j \quad (\text{note } f_j(e) < s_j)$$

Therefore $\eta(e) < 0$ if $e > \Delta + \sum_{j\in M} s_j = \hat{e}$. By the Mean Value Theorem, there exists e such that $\eta(e) = 0$. Let $\bar{e} = \min\{e \geq 0 | \eta(e) = 0\}$; this number exists and is finite.

For $e \in [0, \bar{e})$, $\eta(e) > 0$, hence $|\tilde{\mathbf{s}}(e)| = t_2 - \eta(e) \leq t_2 \leq 1$. The responsive coefficients of all workers except b are the same under $\tilde{\mathbf{s}}(e)$ by part 3 Claim 3.2.2, and $k_b(\tilde{\mathbf{s}}(e)) = s_b + e$ (b has no followers), which is increasing in e.

$$\frac{\partial w(\tilde{\mathbf{s}}(e), \mathcal{H})}{\partial e} = \frac{\partial}{\partial e}\{s_b + e - \frac{1}{2}(s_b + e)^2\} = 1 - (s_b + e) > 1 - |\tilde{\mathbf{s}}(e)| \geq 1 - t_2 \geq 0$$

Therefore, the aggregate welfare $w(\tilde{\mathbf{s}}(e), \mathcal{H})$ is strictly increasing as we increase $e \in [0, \bar{e}]$, while for $e = 0$, $\tilde{s}(0) = \mathbf{s}$, and for $e = \bar{e}$, $|\tilde{\mathbf{s}}(\bar{e})| = t_2 - \eta(\bar{e}) = t_2 - 0 = t_2$. In the end, we have:

$$\begin{aligned}\phi(t_1, \mathcal{H}) &= w(\mathbf{s}, \mathcal{H}) = w(\tilde{\mathbf{s}}(0), \mathcal{H}) \\ &< w(\tilde{\mathbf{s}}(\bar{e}), \mathcal{H}) \leq \max_{\mathbf{s}\geq 0, \sum s_j = t_2} w(\mathbf{s}, \mathcal{H}) = \phi(t_2, \mathcal{H})\end{aligned}$$

which completes the proof of Theorem 1.4.9. ■

Proof of Claim 3.2.2 We can construct these functions step-by-step. For each $e \geq 0, j \in M$, define $f_j(e)$ as the unique solution to the following

$$g(s_j, e + \sum_{l\in F^j} k_l) = g(s_j - f_j(e), \sum_{l\in F^j} k_l)$$

The solution $f_j(e)$ exists and is unique by continuity of g and the fact that $j \in M$ and $s_j > 0$. Also, f_j is continuous by the implicit function theorem.

Last, we need to check the three conditions in Claim 3.2.2. Parts 1 and 2 are obviously true by construction. For part 3, we prove this by induction on the level of members.

If $j \in N^h$, then $k_j(\tilde{\mathbf{s}}(e)) = \tilde{s}_j(e) = s_j$ if $j \neq b$, and $k_b(\tilde{\mathbf{s}}(e)) = \tilde{s}_b(e) = s_b + e$. Suppose part 3 holds for any member on levels higher than K. Suppose $j \in N^{K-1}$. There are two cases.

(1) Suppose $j \in M$. Hence $j \in P^b$, and $b \in F^j$. Then

$$\begin{aligned}k_j(\tilde{\mathbf{s}}(e)) &= g(s_j - f_j(e), \sum_{l\in F^j} k_l(\tilde{\mathbf{s}}(e))) \\ &= g(s_j - f_j(e), k_b(\tilde{\mathbf{s}}(e)) + \sum_{l\in F^j, l\neq b} k_l(\tilde{\mathbf{s}}(e))) \\ &= g(s_j - f_j(e), s_b + e + \sum_{l\in F^j, l\neq b} k_l(\mathbf{s})) \quad \text{(by induction)} \\ &= g(s_j - f_j(e), e + \sum_{l\in F^j} k_l(\mathbf{s})) = g(s_j, e + \sum_{l\in F^j} k_l) \quad \text{(by definition of } f_j(e)\text{)} \\ &= k_j(\mathbf{s})\end{aligned}$$

The fourth equality follows by induction, since the set F^j must lie on a higher level than j.
(2) If $j \notin M$, then $b \notin F^j$. Then

$$k_j(\tilde{\mathbf{s}}(e)) = g(s_j, \sum_{l \in F^j} k_l(\tilde{\mathbf{s}}(e))) \quad = \quad g(s_j, \sum_{l \in F^j} k_l(\mathbf{s})) = k_j(\mathbf{s})$$

By induction, part 3 holds for any member j. ■

This example shows that the result is not as obvious as it appears.

Example 5 *There exists a hierarchy $\mathcal{H}$ and two shares $\mathbf{s}, \mathbf{s}'$ with $\mathbf{s}' \geq \mathbf{s}$, but $w(\mathbf{s}, \mathcal{H}) > w(\mathbf{s}', \mathcal{H})$.*

For the chain $A \to B \to C$ with $s_A = 0.8, s_B = s_C = 0.1$, we have $k_A(\mathbf{s}, C_3) = 1.00782 > 1$. Reduce the share of s_A by 1%, which will reduce k_A to 0.99737 without affecting k_B, k_C. Note $0.99737 - \frac{1}{2}0.99737^2 > 1.00782 - \frac{1}{2}1.00782^2$. The new shares add up to only 99%, but yield higher aggregate welfare. The problem is that the shares are not optimally adjusted as we did in Theorem 1.4.9.

Proof of Lemma 1.4.11

Proof For the chain, the responsive coefficients k_i and k_{i+1} are related by the following:

$$k_i - k_{i+1} = g(x, g(x,t) + t) - g(x,t)$$

where $x = \frac{1}{N} > 0, t = \sum_{j>i+1} k_j \geq 0$. By Lemma 3.1.2, $k_i - k_{i+1}$ lies between $x/2$ and x, in other words, $\frac{1}{2N} < k_i - k_{i+1} < \frac{1}{N}, i = 1, \cdots, N-1$. Taking summations, we have:

$$\frac{N+1-i}{2N} < k_i < \frac{N+1-i}{N} \qquad i = 1, \cdots, N-1$$

which completes the proof of Lemma 1.4.11. ■

Proof of Proposition 1.4.12

Proof Based on the estimates of k_i from Lemma 1.4.11, we have:

$$\begin{aligned} \sum_{k=1}^{N} \left(\frac{N+1-i}{2N} - \frac{1}{2}(\frac{N+1-i}{2N})^2 \right) &\leq w(\mathbf{s}^{eq}, \mathcal{C}_N) = \sum_{k=1}^{N} (k_i - \frac{1}{2}k_i^2) \\ &\leq \sum_{k=1}^{N} \left(\frac{N+1-i}{N} - \frac{1}{2}(\frac{N+1-i}{N})^2 \right) \end{aligned}$$

Simplifying the terms, we get

$$\frac{(1+N)(-1+10N)}{48N} \leq w(\mathbf{s}^{eq}, \mathcal{C}_N) \leq \frac{(1+N)(-1+4N)}{12N}$$

Therefore

$$\frac{5}{24}N + \frac{1}{6} \leq w(\mathbf{s}^{eq}, \mathcal{C}_N) \leq \frac{1}{3}N + \frac{1}{4}$$

For a large team with equal shares, $w(\mathbf{s}^{eq}, \mathcal{C}_N)$ grows at least linearly in N. ■

Proof of Proposition 1.4.13

Proof This proof is quite similar to the proof of Lemma 1.4.11.
For the chain structure, the the responsive coefficients k_i and k_{i+1} are related by the following:

$$\begin{aligned} k_i - k_{i+1} &= g(x, g(y,t)+t) - g(y,t) \\ &= \underbrace{(g(x, g(y,t)+t) - g(y, g(y,t)+t))}_{I} + \underbrace{(g(y, g(y,t)+t) - g(y,t))}_{II} \end{aligned}$$

where $x = s_i, y = s_{i+1}, t = \sum_{j>i+1} k_j$. Note that $x \leq y$ and $t \geq 0$ by assumption. The first term $I = g(x, g(y,t)+t) - g(y, g(y,t)+t) = (x-y)g_x(\zeta, g(y,t)+t)$ for some ζ by the Mean Value Theorem. Note $g_x \geq 1$ and $x - y \leq 0$, so $I \leq (x-y)g_x(x,\zeta) \leq (x-y)$. The second term $II = g(y, g(y,t)+t) - g(y,t) \leq y$ by Lemma 3.1.2. Therefore, $k_i - k_{i+1} = I + II \leq (x-y) + y = x = s_i$. Taking summations, we have

$$k_i \leq \sum_{j \geq i}(k_j - k_{j+1}) \leq \sum_{j \geq i} s_j$$

Note the sum of all shares is one, so $k_i \leq 1$.
If all the shares s_i are positive, then the second term $II < y = s_{i+1}$, which implies $k_i - k_{i+1} < s_i$, therefore $k_i < 1$. ■

Proof of Theorem 1.4.14

Proof The proof consists of two steps.

Step 1: Proof that $\frac{s_i^*}{s_{i+1}^*} \in [0.5, 1)$

Suppose $\mathbf{s}^*$ is optimal for the chain. We want to prove that $\frac{s_i^*}{s_{i+1}^*} \in [0.5, 1)$ for any i. Let $k = s_i^* + s_{i+1}^*, e = s_{i+1}^*$. Then it is easy to see that

$$\frac{s_i^*}{s_{i+1}^*} \in [0.5, 1) \iff \frac{e}{k} = \frac{1}{1 + \frac{s_i^*}{s_{i+1}^*}} \in (\frac{1}{2}, \frac{2}{3}]$$

We prove this is true by contradiction.

Suppose $\frac{e}{k} \in [0, \frac{1}{2}]$, or $e \in [0, \frac{k}{2}]$. Then from Lemma 3.1.4 we have $A'(e) > 0, M'(e) > 0$. So for small $\delta > 0$, the following two conditions hold:

$$A(e+\delta) > A(e) \quad \text{and} \quad M(e+\delta) > M(e) \tag{3.5}$$

Hence we can define the new shares $\hat{\mathbf{s}}$ by moving δ from i to $i+1$ in $\mathbf{s}^*$. Obviously, this will not change the incentives of workers after $i+1$. Also by equation (3.5), the new shares $\hat{\mathbf{s}}$ satisfy:

$$\hat{k}_l = k_l, l > i+1 \tag{3.6}$$
$$\hat{k}_i + \hat{k}_{i+1} > k_i + k_{i+1}, \tag{3.7}$$
$$\hat{k}_i - \frac{1}{2}(\hat{k}_i)^2 + \hat{k}_{i+1} - \frac{1}{2}(\hat{k}_{i+1})^2 > k_i - \frac{1}{2}(k_i)^2 + k_{i+1} - \frac{1}{2}(k_{i+1})^2 \tag{3.8}$$

By induction, we also have $\hat{k}_l > k_l, l < i$.

Because we do not know the range of $\hat{k}_l$, we cannot argue that $w(\hat{\mathbf{s}}, \mathcal{C}_N) > w(\mathbf{s}^*, \mathcal{C}_N)$. Instead, we apply the same trick as before: we reduce the share of $i-1$ by a suitable amount $\epsilon_{i-1} \geq 0$ such that

$$g(s_{i-1} - \epsilon_{i-1}, \hat{k}_i + \hat{k}_{i+1} + \sum_{j>i+1} \hat{k}_j) = g(s_{i-1}, k_i + k_{i+1} + \sum_{j>i+1} k_j)$$

This is always feasible by continuity of g and equation (3.7). Then do the same operations for player $i-2$, $i-3$ through player 1 such that their responsive coefficients for the reduced shares are the same as those with $\mathbf{s}^*$. Call $\tilde{\mathbf{s}}$ the reduced share profile. Then $|\tilde{\mathbf{s}}| = \beta < 1$ by construction. Therefore, $w(\tilde{\mathbf{s}}, \mathcal{C}_N) > w(\mathbf{s}^*, \mathcal{C}_N)$ by equation (3.8). On the other hand,

$$w(\tilde{\mathbf{s}}, \mathcal{C}_N) \leq \phi(\beta, \mathcal{C}_N) \leq \phi(1, \mathcal{C}_N) = w(\mathbf{s}^*, \mathcal{C}_N)$$

So we get a contradiction. Hence it is impossible to have $e \in [0, \frac{k}{2}]$.

Similarly, it is impossible to have $\frac{e}{k} \in (\frac{2}{3}, 1]$ by using Lemma 3.1.4 to get a contradiction.

For the last two workers, we can even show that $s_{N-1} = \frac{1}{2}s_N$. If this is not the case, then we can move part of the share from one worker to the other such that the ratio is 1/2. Notice that in this case we will have $A(2k/3) > A(e)$, and $M(2k/3) > M(e)$ (Part 4 of Lemma 3.1.4). We get a contradiction by using the same method as above.

Step 2: Proof that $k_1 > k_2 > \cdots > k_{N-1} = k_N$

Notice that $s^*_{N-1} = 0.5 s^*_N$, so $k_{N-1} = g(0.5 s^*_N, s^*_N) = s^*_N = k_N$.

For any other pair of players $\{s_i, s_{i+1}\}$, $i \neq N-1$, let $k = s_i + s_{i+1}, \bar{e} = s_{i+1}$. If

$k_i \leq k_{i+1}$, then $\tilde{g} - g = k_{i+1} - k_i \leq 0$, and $\frac{\bar{e}y}{2y+\bar{e}} > 0$ ($y > 0$ in this case because $i+1$ is not the terminal worker). Therefore

$$A'(\bar{e}) < 0, \quad \text{or} \quad \tilde{a}'(\bar{e}) + a'(\bar{e}) < 0$$

by equation 3.3. Also, by monotonicity of s_i in Step 1 and Proposition 1.4.13 we have $k_i \leq k_{i+1} < 1$, therefore $1 > a(\bar{e}) \geq \tilde{a}(\bar{e})$. Then

$$M'(e) = (1-\tilde{a}(e))\tilde{a}'(e)+(1-a(e))a'(e) = (1-\tilde{a}(e))(\tilde{a}'(e)+a'(e))+(\tilde{a}(e)-a(e))a'(e)$$

which is strictly negative at $e = \bar{e}$, since $(1-\tilde{a}(\bar{e}))(\tilde{a}'(\bar{e}) + a'(\bar{e})) < 0$, and $(\tilde{a}(\bar{e}) - a(\bar{e}))a'(\bar{e}) \leq 0$. Therefore we have shown that

$$A'(\bar{e}) < 0, \text{ and } M'(\bar{e}) < 0$$

This means that we can reduce s_{i+1} by a small amout $\delta > 0$, and increase s_i by the same amount δ, such that:

$$A(\bar{e} - \delta) > A(\bar{e}), \text{ and } M(\bar{e} - \delta) > M(\bar{e})$$

The same procedure can be used to get a contradiction. So $k_i > k_{i+1}$ for $\forall i \neq N-1$. Hence, the proof is complete. ■

Proof of Theorem 1.5.1

Proof We want to show that the linear functions given by Theorem 1.5.1 are part of a separating equilibrium with the pessimistic belief assumption. For brevity, let $k_i = k_i(\mathbf{s}, \mathcal{H})$. Now fix a player $i \in N^k$. If $k = h$, i.e., i is a terminal worker, then obviously $k_i = s_i$. Now suppose $k < h$. Let F^i be the set of followers of i. There are two possible deviations for i, upward and downward.

If i deviates downward (we only consider one player deviating, so all other workers on level k are "telling the truth"), then all the players in F^i will use i's effort to update beliefs by the pessimistic belief assumption and choose efforts accordingly. No profitable downward deviating means that:

$$k_i\theta \in \arg\max_{x \leq k_i\theta} s_i\theta \left(x + \sum_{j \in F^i} k_j(\frac{x}{k_i}) \right) - \frac{1}{2}x^2$$

Using the first order condition, this is equivalent to

$$s_i\theta(1 + \sum_{j \in F^i} \frac{k_j}{k_i}) \geq k_i\theta \Longleftrightarrow k_i \geq g(s_i, \sum_{j \in F^i} k_j)$$

Clearly $k_i = g(s_i, \sum_{j\in F^i} k_j)$ satisfies the condition above (this is actually an equality in this case).
If i deviates upward, the situation is a little bit different, as not all the followers will "listen to" i's effort. Let

$$F_c^i = \{j \in F^i | \text{there is no path from } l \text{ to } j \text{ for } l \in N^k, l \neq i\}.$$

These workers will follow i because they cannot detect i's upward deviation as i is the only source of information for them. The workers in $F^i \backslash F_c^i$ will not be affected by i's upward deviation as they get at least one other signal saying that the state is θ. Thus, no profitable upward deviation means that:

$$k_i\theta \in \arg\max_{x\geq k_i\theta} s_i\theta \left(x + \sum_{j\in F_c^i} k_j(\frac{x}{k_i}) \right) - \frac{1}{2}x^2$$

Using the first order condition, this is equivalent to

$$s_i\theta(1+\sum_{j\in F_c^i}\frac{k_j}{k_i}) \leq k_i\theta \Longleftrightarrow k_i \geq g(s_i, \sum_{j\in F_c^i} k_j)$$

Obviously, $k_i = g(s_i, \sum_{j\in F^i} k_j) \geq g(s_i, \sum_{j\in F_c^i} k_j)$ satisfies this condition, as $F_c^i \subset F^i$.

Combing these two results, we have shown that $k_i\theta$ is i's best response given all the other players' best responses, thus we have verified that it is part of a separating equilibrium supported by pessimistic beliefs. ■

Remark 7 *As we have seen implicitly in the above proof, any number in the interval* $[g(s_i, \sum_{j\in F_c^i} k_j), g(s_i, \sum_{j\in F^i} k_j)]$ *is a possible choice for i's equilibrium responsive coefficient. We choose the largest one in that interval. The same thing happens in the V structure (Section 2). I conjecture that the equilibrium efforts characterized in Theorem 1.5.1 are the upper bounds of equilibrium efforts among all separating equilibria supported by pessimistic beliefs as I show for the V structure in section 2.*

Proof of Theorem 1.5.3

Proof Suppose $\mathbf{s}$ is optimal for $\phi(t, \mathcal{H})$, i.e., $\phi(t, \mathcal{H}) = w(\mathbf{s}, \mathcal{H})$. We claim that there exists a contract $\mathbf{s}'$ such that $w(\mathbf{s}, \mathcal{H}) = w(\mathbf{s}', \mathcal{H} + ij)$ and $\mathbf{s}' \leq \mathbf{s}$. Then the proposition follows directly from this claim and Theorem 4.9.
To show the claim, first we claim that there exists a $\delta \geq 0$ such that:

$$g(s_i, \sum_{t\in F^i} k_t(\mathbf{s}, \mathcal{H})) = g(s_i - \delta, \sum_{t\in F_a^i} k_t(\mathbf{s}, \mathcal{H}))$$

Here F_a^i is the set of i's followers in $\mathcal{H} + ij$, which is larger than F^i because a link from i to j was added. The existence of δ follows from continuity of g. By induction, we can keep weakly reducing all players on the top of l while keeping their incentives the same under the two hierarchies. As everyone has weakly more followers after adding the link, we can always do that. In the end, let $\mathbf{s}'$ be the resulting new contract. Then $\mathbf{s}' \leq \mathbf{s}$. Also, by construction:

$$k_l(\mathbf{s}, \mathcal{H}) = k_l(\mathbf{s}', \mathcal{H} + ij), \forall l \in \mathcal{N}.$$

So $w(\mathbf{s}, \mathcal{H}) = w(\mathbf{s}', \mathcal{H} + ij)$. That establishes the claim. ■

Proof of Proposition 1.5.5

Proof Let $t_1, t_2, \cdots, t_{n_k}$ be the shares of workers on the same level, say N^k, in an optimal contract $\mathbf{s}$. Let $y = \sum_{l>k} \sum_{j \in N^l} k_j$. Then by Theorem 4.3, the responsive coefficients of these n_k members are given by $g(t_i, y)$, and the contribution to welfare by these workers is

$$G(t_1, \cdots, t_{n_k}) := \sum_{i=1}^{n_k} \left\{ g(t_i, y) - \frac{1}{2} g(t_i, y)^2 \right\}$$

Define $F(t_1, \cdots, t_{n_k}) := \sum_{i=1}^{n_k} \{g(t_i, y)\}$ as the sum of the responsive coefficients. We prove the proposition by contradiction.
If $t_m \neq t_n$ for $m \neq n$, then let $s = \frac{\sum_{i=1}^{n_k} t_i}{n_k}$ be the new equalized share for members in N^k. We claim that

$$G(t_1, \cdots, t_{n_k}) < G(s, \cdots, s) \text{ and } F(t_1, \cdots, t_{n_k}) < F(s, \cdots, s) \tag{3.9}$$

Note $g(t, y)$ is strictly concave in t (Lemma 3.1.1), so

$$F(t_1, \cdots, t_{n_k}) = n_k \sum_{i=1}^{n_k} \left\{ \frac{1}{n_k} g(t_i, y) \right\} < n_k \left\{ g(\frac{\sum_{i=1}^{n_k} t_i}{n_k}, y) \right\} = n_k g(s, y) = F(s, \cdots, s)$$

by Jensen's inequality.
Similarly, $g(t, y) - \frac{1}{2} g(t, y)^2 = (1 - \frac{1}{2} t) g(t, y) - \frac{1}{2} t y$ is concave in t because

$$(g(t, y) - \frac{1}{2} g(t, y)^2)'' = (1 - \frac{1}{2} t) g_{xx}(t, y) - g_x(t, y) < 0$$

Here $t < 1, 1 - \frac{1}{2} t > 0, g_{xx} < 0, g_x > 0$. By the same logic, we can show $G(t_1, \cdots, t_{n_k}) < G(s, \cdots, s)$.
For Equation (3.9), we can get a similar contradiction as shown in the proof of Theorem 1.4.14. We do not repeat the argument here. ■

Proof of Proposition 1.5.6

Proof Let $\mathbf{s}$ be the optimal share profile for $\phi(t,\mathcal{H})$. We compare the coefficients k_i for $\mathbf{s}$ under $\mathcal{H}$ and $\mathcal{H}'$. The incentives of workers in $N^{k+1},\cdots,N^h$ and $N^k\backslash\{i\}$ are obviously the same in both cases. Worker i's incentive is different because i has more followers, so

$$k_i'(\mathbf{s},\mathcal{H}') = g\left(s_i, (\sum_{j\in N^k\backslash\{i\},} k_l(\mathbf{s},\mathcal{H}')) + y\right) > k_i(\mathbf{s},\mathcal{H}') = g(s_i,y)$$

where $y = \sum_{l=k+1}^{h}\sum_{j\in N^l} k_j(\mathbf{s},\mathcal{H})$. Therefore, we can reduce the share of s_i by a small amount $\delta > 0$, such that

$$g\left(s_i-\delta, (\sum_{j\in N^k\backslash\{i\},} k_l(\mathbf{s},\mathcal{H}')) + y\right) = k_i(\mathbf{s},\mathcal{H}') = g(s_i,y)$$

With this reduction, the incentives for workers in N^k are also the same as before. Then by induction, the responsive coefficients for workers in $N^{k-1},\cdots,N^1$ are also the same. Therefore, we can find a share profile for $\mathcal{H}'$ that uses less total share (adds up to $1-\delta$) and yields the same welfare. Extra share is welfare improving by Theorem 4.9, hence $\phi(t,\mathcal{H}) < \phi(t,\mathcal{H}')$. ■

Proof of Equation 1.11 and Proposition 1.6.2

Proof: Once we have equation 1.11, Proposition 1.6.2 follows from a similar argument. So, it suffices to show equation 1.11 holds for any hierarchy in $\mathcal{M}^s(N,K,1)$.

Suppose $\mathcal{H} \in \mathcal{M}^s(N,K,1)$. Take any middle manger, say M with his q followers. Suppose shares of the middle manager and his followers are $u, v_1, v_2, \cdots, v_q$. Then we claim that:

$$(\dagger) \qquad g(u,\sum_{i=1}^{q} v_i) + \sum_{i=1}^{q} v_i \le \frac{4}{3}(u + \sum_{i=1}^{q} v_i)$$

Taking summation over all the middle level workers, then the right hand side will be $\frac{4}{3}$ times the sum of shares of $2p$ followers, which is exactly $\frac{4}{3}(1-l)$. The left hand side will be the sum of the responsive coefficients of those $2p$ workers. That is exactly equation 1.11. Inequality (†) follows from

$$g(x,y)+y \le \max_{t\in[0,x+y]} g(x+y-t,t)+t = g(x+y-t,t)+t|_{t=\frac{2}{3}(x+y)} = \frac{4}{3}(x+y),$$

while the equality in the middle follows from part 4 of Lemma 3.1.4.

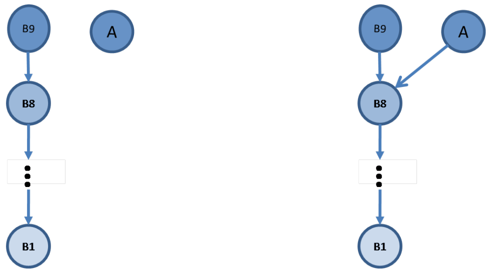
If any middle level worker has more than one follower, then either equation 1.11 is strict, or equation 1.12 is strict. In this case we have a strict welfare comparison, which shows that $\mathcal{H}^1$ is the most efficient hierarchy in $\mathcal{M}^s(N, K, 1)$. ■

Remark 8 *The proof also shows that to achieve the highest welfare under* $\mathcal{H}^1$*, we must assign* $m = \frac{1-l}{3p}$ *to each manager and* $f = \frac{2(1-l)}{3p}$ *to each terminal worker, and* l *to the leader for some* $l \in (0, 1)$*. The only unknown variable is the number* l*. The optimal* l *can be determined by solving the corresponding welfare maximization program.*

3.3 Additional Materials

Adding Links is Not Always Welfare Improving

Take a hierarchy as shown in figure 3.1a, in which A and B_9 are on level 1, and B_j are on level $10 - j$ for each j. Their shares are: $s_A = 91\%$ and $s_{B_i} = 1\%, i = 1, \cdots, 9$. The responsive coefficients k_{B_i} are: $\{0.01, 0.016, 0.022, 0.027, 0.033, 0.038, 0.043, 0.049, 0.054\}$, and $k_A = s_A = 0.9$. If one additional link is added from A to B_8 as shown in Figure 3.1b, then A's responsive coefficient under the new structure is $\tilde{k}_A = g(0.9, \sum_{i=1}^{8} k_{B_i}) = 1.10688$, which is further away from $k^{FB} = 1$ than $k_A = 0.9$. Moreover, this transformation only affects the incentive for A; the equilibrium efforts for other members are not affected. Therefore the aggregate welfare of the new hierarchy with this additional link is actually lower than the original one. This is not inconsistent with Theorem 5.3 because we have not adjusted the shares optimally here.



(a) old hierarchy

(b) new hierarchy

Figure 3.1: A counterexample: adding links is not always welfare improving.

More About Belief Functions in V Structure

We explore more possible belief functions in the V structure in this subsection. Unlike the analysis with the chain structure, conditions $(P1'), (P2'), (B')$ are not enough to uniquely pin down the equilibrium efforts in this case. In general, different belief functions can support different equilibrium efforts, and one belief function can support multiple equilibria.

Definition 8 *Given a belief function β satisfying B', we say e_1, e_2* **are supported** *by β if e_1, e_2 and β satisfy conditions $S, P1', P2'$.*

To find equilibria, it suffices to find functions e_1, e_2 that can be supported by a given β. In the following examples, we list some special belief functions and find the effort functions that can be supported by each.

The first one is the one used in the paper.

1. *Pessimistic belief*

$$\beta^p(\theta_1, \theta_2) = \min(\theta_1, \theta_2) \tag{3.10}$$

In this case, using FOC, we can rewrite $P1', P2'$ as:

$$\forall \theta,\ \frac{\theta}{3} \leq e_i(\theta) \leq \frac{\theta}{3}\left(1 + \frac{1}{3e_i'(\theta)}\right), \quad i = 1, 2. \tag{3.11}$$

There are multiple solutions to equation 3.11. For example, $e_i(\theta) = k_i\theta$, for any $k_i \in [\frac{1}{3}, \frac{1+\sqrt{5}}{6}], i = 1, 2$. Nevertheless, the solutions to equation 3.11 are bounded in the following sense.

Lemma 3.3.1 *If e_i is monotonic and satisfies equation 3.11, then*

$$\forall \theta,\ \underline{e}(\theta) \leq e_i(\theta) \leq \bar{e}(\theta)$$

Here, $\bar{e}(\theta) := g(\frac{1}{3}, \frac{1}{3})\theta$ and $\underline{e}(\theta) := \frac{\theta}{3}$ are defined in section 2. So, $\bar{e}(\theta)$ and $\underline{e}(\theta)$ are the upper and lower bounds for the solutions to equation 3.11. Moreover, both $\bar{e}(\theta)$ and $\underline{e}(\theta)$ satisfy equation 3.11.

2. *Trigger belief*

$$\beta^t(\theta_1, \theta_2) = \begin{cases} \theta_1 & \text{if } \theta_1 = \theta_2 \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

In this case both leaders have the same information about the state. If any leader deviates, F will detect it and choose zero effort under this belief assumption. Moreover, $P1', P2'$ are equivalent to:

$$\left(e_i(\theta) - \frac{1}{3}\theta\right)^2 \leq \frac{2}{9}\theta^2, \quad i = 1, 2. \tag{3.13}$$

In particular, $e_i(\theta) = k_i\theta$, for any $k_i \in (0, \frac{1+\sqrt{2}}{3}], i = 1, 2$ will satisfy these conditions. Not all solutions to the above equations are linear.

3. *Weighted belief*
$$\beta^{w1,w2}(\theta_1, \theta_2) = w_1\theta_1 + w_2\theta_2 \tag{3.14}$$
for $0 \leq w_i \leq 1, w_1 + w_2 = 1$. Under this belief, $P1', P2'$ can be written as:
$$\frac{\theta}{3}\left(1 + \frac{w_i}{3e_i'(\theta)}\right) = e_i(\theta), \quad i = 1, 2.$$
The solutions to this differential equation are linear in θ, and given by
$$e_i(\theta|w_i) = g(\frac{1}{3}, \frac{w_i}{3})\theta, \; i = 1, 2. \tag{3.15}$$
Note that $\underline{e}$ and $\bar{e}$ are special cases with $w_i = 0$, and $w_i = 1$. But, this belief function cannot support $e_i = \underline{e}, \forall i$ or $e_i = \bar{e}, \forall i$, because otherwise $w_1 + w_2 \neq 1$.

   Not every specification of β is consistent with equilibrium, as the next example shows.

4. *Optimistic belief*
$$\beta^o(\theta_1, \theta_2) = \max(\theta_1, \theta_2) \tag{3.16}$$
In this case, $P1', P2'$ can be written as:
$$\frac{\theta}{3} \geq e_i(\theta) \geq \frac{\theta}{3}\left(1 + \frac{1}{3e_i'(\theta)}\right), \quad i = 1, 2.$$
Note that $\frac{\theta}{3} < \frac{\theta}{3}\left(1 + \frac{1}{3e_i'(\theta)}\right)$. These two inequalities are inconsistent, hence there is no separating equilibrium with optimistic belief.

The trigger belief β^t can generate not only very efficient outcomes, for example $e_i(\theta) = \frac{1+\sqrt{2}}{3}\theta, i = 1, 2$, but also very inefficient ones, for example $e_i(\theta) = \epsilon\theta, i = 1, 2$ for arbitrarily small $\epsilon > 0$. A unsatisfactory fact about β^t is that it has jumps and it is not monotonic in θ_1, θ_2. For β^p, and $\beta^{w1,w2}$, the equilibrium efforts of both leaders are at least $\frac{\theta}{3}$.

Sometimes, the following restrictions are natural to assume.

(BDM) $\quad$ β is continuous, differentiable and monotonic in θ_1, θ_2

With this assumption (BDM), $P1'$ and $P2'$ can be replaced by the corresponding FOCs.
$$\frac{\theta}{3}\left(1 + \frac{\beta_i(\theta, \theta)}{3e_i'(\theta)}\right) = e_i(\theta), \; i = 1, 2 \tag{3.17}$$

where $\beta_i = \frac{\partial \beta}{\partial \theta_i}$. For any β satisfying (BDM), we can find an equilibrium by solving the differential equations (3.17). Although the solutions are in general different for different βs, nevertheless, we claim:

Claim 3.3.2 *All solutions to equation 3.17 satisfy equation* 3.11.

Proof Differentiating both sides of $\beta(\theta, \theta) = \theta$ gives $\beta_1(\theta, \theta) + \beta_2(\theta, \theta) = 1$. Note β is monotonic, so $\beta_i(\theta, \theta) \geq 0$. Therefore $0 \leq \beta_i(\theta, \theta) \leq 1$. By equation 3.17,

$$\frac{\theta}{3} = \frac{\theta}{3}\left(1 + \frac{0}{3e_i'(\theta)}\right) \leq e_i(\theta) = \frac{\theta}{3}\left(1 + \frac{\beta_i(\theta, \theta)}{3e_i'(\theta)}\right) \leq \frac{\theta}{3}\left(1 + \frac{1}{3e_i'(\theta)}\right)$$

which is exactly equation 3.11. ■

This immediately yields the following result.

Proposition 3.3.3 *If $\{e_i, i = 1, 2\}$ can be supported by a belief function β satisfying (BDM), then $\{e_i, i = 1, 2\}$ can also supported by pessimistic belief β^p.*

Of course, β is part of the equilibrium. In general, we cannot impose assumptions on the endogenous belief functions. We believe that (BDM) is satisfied by a large class of belief functions, although β^t, β^p violate (BDM). Moreover, we can show that Proposition 3.3.3 also holds for β satisfying a weaker differentiability condition than (BDM): existence of left and right derivatives (not necessarily equal). In particular, pessimistic belief β^p satisfies this weaker condition. In the text we argued that assuming β^p as the out-of-equilibrium belief of F makes some sense. In section 2, we found the upper and lower bounds on the corresponding welfare with β^p. By Proposition 3.3.3, these upper and lower bounds hold for any belief satisfying (BDM) or a weaker differentiability condition.[3]

Proof of Lemma 3.3.1

Proof It is easy to see that $e_i(\theta) \geq \frac{\theta}{3}$. If $e_i(\theta) \geq g(\frac{1}{3}, \frac{1}{3})\theta$, then equation 3.11 implies $g(\frac{1}{3}, \frac{1}{3}) \leq \frac{e_i(\theta)}{\theta} \leq \frac{1}{3}\left(1 + \frac{1}{3e_i'(\theta)}\right)$. Equivalently, $e_i'(\theta) \leq g(\frac{1}{3}, \frac{1}{3})$. Let $f(\theta) = e_i(\theta) - g(\frac{1}{3}, \frac{1}{3})\theta$. So $f'(\theta) \leq 0$ whenever $f(\theta) \geq 0$. Also $e_i(0) = 0$, so $f(0) = 0$. Lemma 3.1.5 shows that $f(\theta) \leq 0, \forall\theta$, or equivalently $e_i(\theta) \leq g(\frac{1}{3}, \frac{1}{3})\theta, \forall\theta$. ■

[3] However, for trigger belief β^t, we can support $e_i = \frac{1+\sqrt{2}}{3}\theta, i = 1, 2$. The corresponding welfare is $\frac{(11+8\sqrt{2})}{18}\theta^2 \approx 1.23965\,\theta^2$, which is higher than W^S.

Theorem 2.3.1, Proof of Existence

To show: for each Λ, there exists a set S with the property that:

1. $\nu(S) = \Lambda$,

2. There exists a finite number α such that $g \geq \alpha$ on S, and $g \leq \alpha$ on the complement.

To this end, for each $a \in (-\infty, \infty)$, define $t(a) = \nu\{g \leq a\}$, and define right and left limits:
$t_-(a) := \lim_{n\to\infty} t(a - \frac{1}{n}) = \lim_{n\to\infty} \nu\{g \leq a - \frac{1}{n}\} = \nu(\{g < a\})$
$t_+(a) := \lim_{n\to\infty} t(a + \frac{1}{n}) = \lim_{n\to\infty} \nu\{g \leq a + \frac{1}{n}\} = \nu(\{g \leq a\}) = t(a)$.

Because t is monotonic increasing, both the left limit $t_-(a)$ and right limit $t_+(a)$ exist. Moreover, t is right continuous, but may have jumps. Let $t_+(a) - t_-(a) = \nu(\{g = a\}) \geq 0$ be the jump at a.

Let $\lambda = 1 - \Lambda$, hence $\lambda \in (0, 1)$. Define $\alpha = \inf\{a \mid t(a) \geq \lambda\}$. Then α is finite and $t_-(\alpha) \leq \lambda \leq t_+(\alpha) = t(\alpha)$. The jump at α is $t(\alpha) - t_-(\alpha)$. There are two cases:

1. If there is zero jump at α, choose $S = \{g > \alpha\}$. Then $\nu(S) = 1 - t(\alpha) = 1 - \lambda = \Lambda$. Clearly $g \leq \alpha$ on the complement of S.

2. If there is a positive jump at α, $\nu(\{g = \alpha\} = t(\alpha) - t_-(\alpha) > 0$. Then $0 \leq \lambda - t_-(\alpha) \leq t(\alpha) - t_-(\alpha)$. Since ν is atomless, there exists a subset $P_1 \subset \{g = \alpha\}$ with measure $\nu(P_1) = \lambda - t_-(\alpha)$. Let $P_2 = \{g = \alpha\} \backslash P_1$, $P^+ = \{g > \alpha\}$, $P^- = \{g < \alpha\}$, then (P^+, P^-, P_1, P_2) is a partitioning of the whole space. Choose $S = P^+ \cup P_2$. Then the complement of S is $P^- \cup P_1$, so $g \leq \alpha$ on the complement, and $\nu(\{g < \alpha\}) + \nu(P_1) = t_-(\alpha) + \lambda - t_-(\alpha) = \lambda$. Therefore and $\nu(S) = 1 - \lambda = \Lambda$ and $g \geq \alpha$ on S. ■

Bibliography

[1] James Andreoni, *Leadership giving in charitable fund-raising*, Journal of Public Economic Theory **8** (2006), no. 1, 1–22.

[2] S. Athey, *Monotone comparative statics under uncertainty*, The Quarterly Journal of Economics **117** (2002), no. 1, 187–223.

[3] Venkatesh Bala and Sanjeev Goyal, *Learning from neighbours*, The Review of Economic Studies **65** (1998), no. 3, 595–621.

[4] A.V. Banerjee, *A simple model of herd behavior*, The Quarterly Journal of Economics **107** (1992), no. 3, 797.

[5] S. Bikhchandani, D. Hirshleifer, and I. Welch, *A theory of fads, fashion, custom, and cultural change as informational cascades*, Journal of political Economy (1992), 992–1026.

[6] P. Bolton, M.K. Brunnermeier, and L. Veldkamp, *Leadership, coordination and mission-driven management*, 2008.

[7] ______, *Economists' perspectives on leadership*, Handbook of Leadership Theory and Practice, Harvard Business School Press, 2010, p. 239.

[8] P. Bolton and M. Dewatripont, *The firm as a communication network*, The Quarterly Journal of Economics **109** (1994), no. 4, 809.

[9] G.A. Calvo and S. Wellisz, *Supervision, loss of control, and the optimum size of the firm*, The Journal of Political Economy (1978), 943–952.

[10] ______, *Hierarchy, ability, and income distribution*, The Journal of Political Economy (1979), 991–1010.

[11] I.K. Cho and D.M. Kreps, *Signaling games and stable equilibria*, The Quarterly Journal of Economics **102** (1987), no. 2, 179.

[12] L. Garicano, *Hierarchies and the organization of knowledge in production*, Journal of Political Economy **108** (2000), no. 5, 874–904.

[13] B.E. Hermalin, *Toward an economic theory of leadership: Leading by example*, The American Economic Review **88** (1998), no. 5, 1188–1206.

[14] ______, *Leadership and corporate culture*, forthcoming Handbook of Organization (2007).

[15] ______, *Leading for the long term*, Journal of Economic Behavior & Organization **62** (2007), no. 1, 1–19.

[16] B. Holmstrom, *Moral hazard in teams*, The Bell Journal of Economics **13** (1982), no. 2, 324–340.

[17] M.O. Jackson, *Social and economic networks*, Princeton Univ Press, 2008.

[18] M. Komai and M. Stegeman, *Leadership based on asymmetric information*, The RAND Journal of Economics **41** (2010), no. 1, 35–63.

[19] M. Komai, M. Stegeman, and B.E. Hermalin, *Leadership and information*, The American Economic Review **97** (2007), no. 3, 944–947.

[20] G.J. Mailath, *Incentive compatibility in signaling games with a continuum of types*, Econometrica (1987), 1349–1365.

[21] T. Marschak and S. Reichelstein, *Communication requirements for individual agents in networks and hierarchies*, 1995.

[22] ______, *Network mechanisms, informational efficiency, and hierarchies*, Journal of Economic Theory **79** (1998), no. 1, 106–141.

[23] P. Milgrom and C. Shannon, *Monotone comparative statics*, Econometrica (1994), 157–180.

[24] P.R. Milgrom, *Good news and bad news: Representation theorems and applications*, The Bell Journal of Economics (1981), 380–391.

[25] P.R. Milgrom and R.J. Weber, *A theory of auctions and competitive bidding*, Econometrica: Journal of the Econometric Society (1982), 1089–1122.

[26] J. Potters, M. Sefton, and L. Vesterlund, *Why announce leadership contributions?: An experimental study of the signaling and reciprocity hypotheses*, Tilburg University, 2001.

[27] ———, *After you–endogenous sequencing in voluntary contribution games*, Journal of Public Economics **89** (2005), no. 8, 1399–1419.

[28] A. Prat, *Hierarchies of processors with endogenous capacity*, Journal of Economic Theory **77** (1997), no. 1, 214–222.

[29] Y. Qian, *Incentives and loss of control in an optimal hierarchy*, The Review of Economic Studies **61** (1994), no. 3, 527.

[30] M. Rothschild and J.E. Stiglitz, *Increasing risk: I. a definition*, Journal of Economic theory **2** (1970), no. 3, 225–243.

[31] S. Scotchmer, *Risk taking and gender in hierarchies*, Theoretical Economics (2008).

[32] M. Shaked and J.G. Shanthikumar, *Stochastic orders*, Springer Verlag, 2007.

[33] D.M. Topkis, *Minimizing a submodular function on a lattice*, Operations Research (1978), 305–321.

[34] T. Van Zandt, *Real-time decentralized information processing as a model of organizations with boundedly rational agents*, The Review of Economic Studies **66** (1999), no. 3, 633.

[35] L. Vesterlund, *The informational value of sequential fundraising*, Journal of Public Economics **87** (2003), no. 3-4, 627–657.