Design for Channel Coordination

Abstract: Design of network systems is an engineering effort of providing communication services. In this paper, we investigate a strategic implication of design decision on channel coordination. In our model, the system is designed by an upstream monopoly vendor and used by downstream retailers who provide service to end users. We show that the joint presence of demand uncertainty and intra-channel competition can lead to the so-called "Bertrand Supertrap", i.e., a seemingly inferior design that limits firms' ability to expand capacity is actually desirable for maximizing channel profit. We derive a threshold condition for such design to be chosen and also discuss impacts of the design on social welfare and consumer surplus.

1. Introduction

Channel coordination is of vital importance when successive stages of a value chain are operated by separate companies. Aligning competing interests of different firms requires not only a wise choice of channel structure (McGuire and Staelin (2008), Coughlan and Wernerfelt (1989), Trivedi (1998), Chiang et al. (2003)) but also a careful design of vertical constraints that govern the relationship between upstream manufacturer(s) and downstream retailer(s) (Katz (1989), Cachon (2003)). For the latter purpose, many contracting mechanisms have been developed for various situations. The channel can be a dyadic monopoly (Jeuland and Shugan (2008), Moorthy (1987)) or contain one or more competitive stages (Ingene and Parry (1995), Ingene and Parry (2000), Geylani et al. (2007)). Relationships between channel members can be informal and simplistic or strict and complex (Lal (1990)). Firms can be perfectly rational or exhibit some behavioral biases (Ho and Zhang (2008)). The objective can be maximizing profit only or involve other considerations as well (Chu and Desai (1995), Cui et al. (2007)). However, “in a real richer environment”, firms seem to prefer simpler forms of contracts than “the theory often predicts” (Cui et al. (2007)). Moreover, lab experiments also showed that some major features of coordinating contracts, such as buy-back option and revenue sharing, though beneficial, are not as effective as expected by the
theory (Katok and Wu (2009)). These findings necessitate considerations of unconventional ways to help with channel coordination. In this paper, we investigate an indirect mechanism that can be applied to the communications industry: the design of network systems.

Deploying a new generation of network systems is an expensive endeavor that can be undertaken only sporadically. Capacity expansion during interim periods is a major concern of system design. We focus on two opposite approaches: the open-system design, which installs a minimum system initially but allows gradual and limitless capacity increase over time; and the capacitated design, which builds into the system all hardware components for delivering full capacity at the very beginning. Both designs can be implemented by adopting appropriate products. The open-system design is exemplified by the use of Wavelength Division Multiplexers (WDM) in optical networks. The WDM technology allows many transmission devices to use different wavelengths to share the same strand of fiber. Each fiber cable contains many strands and each strand can support hundreds of wavelengths, so by adding transmission devices incrementally, capacity can be expanded continuously to accommodate unlimited growth of demands and the cost of deploying capacity can be spread out over time. On the other hand, the capacitated design can be observed in wireless networks that use software-enabled base stations (e.g., CDMA micro cell base stations). These stations are equipped with all necessary hardware at time of initial installation. The cost of capacity expansion is reduced because it only requires an engineer to use some controlling software to activate pre-installed components instead of visiting station sites to install new equipment. However, the approach requires more investment up-front and is constrained by the amount of capacity installed at the first time.

How to design a system intelligently has been a subject of operations research (Manne (1967), Freidenfields (1981), Luss (1982), Singh et al. (2009) and Andrade et al. (2005)). Typically, the problem is formulated as a multi-period cost minimization model and its solution recommends what products to install and in which period. Economy of scale is the main reason for choosing the capacitated design, which is derived by replacing many small-scale hardware installations with a few big ones, as is illustrated by the above example of micro cell. On the other hand, by
deploying the system incrementally, the open-system design has the advantage of saving time cost of capital, exploiting future reductions of equipment cost, and more importantly, allowing more flexible adjustment of capacity level to match realized demand. Depending on input parameters, the optimal road map may recommend sporadic deployments of large systems with fixed capacity (capacitated design), frequent placements of products of incremental capacity (open-system design), or a mixture of the two.

In this paper, we examine the same design question in a channel context. Instead of assuming a single firm builds a communication system and uses it to provide end user services, we split the two functions into separate companies. The system is built by a monopoly manufacturer and used by many competing service providers. By buying a system and associated capacity components, a service provider effectively acquires a bundle of network capacity that it can sell to end users on retail basis. We capture this situation by a channel model in which the manufacturer is referred as the vendor and service providers are referred as retailers. We consider a two-stage model where retailers purchase the initial system in the first stage and acquire capacity components as needed in the second one. The transaction is carried out under a two-part tariff, which is committed in the first period when end user demand is uncertain. Demand realization is observed in the second period when retailers determine the amount of capacity to buy.

The new formulation leads to the discovery of an interesting situation, which we identify with the so-called “Bertrand Supertrap”. As discussed in (Cabral and Villas-Boas (2005)), a Bertrand Supertrap represents a situation in which a business option associated with desirable properties (e.g., improving cost, achieving economy of scope) is actually a poor solution to be implemented in a competitive environment. Conversely, a seemly mediocre or even inferior solution may be more profitable, not by its own merit, but because its influence on competition. Besides examples given in (Cabral and Villas-Boas (2005)), there are other instances in literature that demonstrate such traps. In the choice of channel structure, “a producer may want to place one or more levels of intermediary between itself and the marketplace even when the producer is capable of carrying out the selling function with the same efficiency as the intermediary.” (McGuire and Staelin (2008)).
In the decision about outsourcing, two competing service providers strictly prefer to delegate their operations to a common supplier “even if the supplier’s technology is no better than the firms’ technology and the supplier is required to establish dedicated capacity (so the supplier’s scale can be no greater than either firm’s scale).” (Cachon and Harker (2002)). We show that a similar trap emerges in the design of network systems because of the joint presence of demand uncertainty and intra-channel competition. Specifically, when the benefit of economy of scale is weak or absent (the latter is assumed in the paper), the capacitated design has no cost advantage and thus can be trivially ruled out by a cost minimization model. However, in a channel model, such design can be desirable because it enhances coordination. We derive a threshold condition under which the coordination benefit is strong enough to render the design a better choice. We also demonstrate impacts of the design on society, consumers, and allocation of profit risk in the channel.

We introduce our modeling assumptions in Section 2, present our main results in Section 3, discuss relevant issues in Section 4, and conclude the paper in Section 5.

2. Problem Setup

We consider a two-stage channel model. There is one monopoly vendor and $N$ retailers. The vendor charges each retailer a two-part tariff. The fixed fee, $F$, is paid for the initial system installed in the first period and the (per-unit) wholesale price, $r$, applies to the purchase of additional capacity in the second period. To avoid the seller’s opportunism problem (McAfee and Schwartz (1994), McAfee and Schwartz (2004)), the vendor commits to a uniform, nondiscriminatory contract that ensures participation of all retailers. The vendor has the bargaining power to extract all surplus and thus is interested in maximizing the expected channel profit.

Retailers offer a homogeneous service and engaged in an intra-channel Cournot competition. We assume a linear demand function

$$p = \theta - \sum_{i=1}^{N} q_i,$$  

where $q_i$ is the amount of capacity that retailer $i$ ($i = 1, 2, ..., N$) supplies, $p$ is the retail price, and parameter $\theta$ characterizes the strength of end user demand. We model demand uncertainty by
letting $\theta$ to be a random variable over a finite support $[\underline{\theta}, \bar{\theta}]$, where $\underline{\theta}$ and $\bar{\theta}$ represent the worst and best demand scenarios respectively. In the first period, all firms have a common view about the distribution of $\theta$, given by the cumulative distribution function $F(\theta)$ and the density distribution function $f(\theta)$. The realization of $\theta$ is observed in the second period based on which retailers decide how much capacity they want to acquire.

3. A Comparison of Two Designs

We first consider the open-system design. Without loss of generality, we normalize the fixed cost of building and deploying the initial system to 0. Assume in the second period, the vendor provides capacity at a constant per-unit cost $C$ and sell at the wholesale price $r$ to retailers who are engaged in the following Cournot competition:

$$\max_{q_i} \left\{ (p - r)q_i = \left( \theta - \sum_{j=1}^{N} q_j - r \right) q_i \right\} \quad i = 1, \ldots, N. \tag{2}$$

The equilibrium outcome can be derived by a standard approach. Denote $\max(0, x)$ by $x^+$. The output of each retailer is

$$q^o(r, \theta) \equiv \left( \theta - r \right)^+ / (N + 1), \tag{3}$$

the equilibrium retail price is

$$p^o(r, \theta) = \theta - N \left( \theta - r \right)^+ / (N + 1), \tag{4}$$

and the equilibrium profit of each retailer is

$$\pi^o(r, \theta) = \left[ p^o(r, \theta) - r \right] q^o(r, \theta). \tag{5}$$

(Note that if $r \geq \theta$, no consumer is willing to pay more than the wholesale price, so no retailer makes any sale.) The expected channel profit is

$$\Pi^o_o(r) = \mathbb{E}[\pi^o(r, \theta) + (C - r)q^o(r, \theta)] = \mathbb{E} \left[ \left( \theta - \frac{N(\theta - r)^+}{N + 1} - C \right) \left( \frac{\theta - r}{N + 1} \right)^+ \right], \tag{6}$$

which is optimized by the vendor’s choice of $r$, i.e.,

$$\Pi^*_o = N \max_{r \geq 0} \{ \Pi^o_o(r) \}. \tag{7}$$
In a perfectly-coordinated channel, the total output would be optimized for each realization of \( \theta \). Denote the expected profit in this ideal situation by \( \Pi^{**} \). The following proposition gives the difference between \( \Pi^{**} \) and \( \Pi^o \).

**Proposition 1.** If
\[
2N\theta \geq (N-1)\mathbb{E}[\theta] + (N + 1)C
\]
and the hazard rate \( f(\theta)/\bar{F}(\theta) \) does not decreases with \( \theta \). Then
\[
\Pi^{**} - \Pi^o = \frac{(N-1)^2}{4(N+1)} \sigma^2(\theta),
\]
where \( \sigma(\theta) \) is the standard deviation of \( \theta \).

The proposition shows that a strictly positive profit gap exists except in two cases: when there is only one retailer \( (N = 1) \) or when there is no uncertainty about future demand \( (\sigma(\theta) = 0) \). The first case is well-understood: a dyadic monopoly channel can be perfectly coordinated if the wholesale price is set at the marginal cost (i.e., \( r = c \)), a strategy that is not contingent upon \( \theta \). When there are many retailers, the vendor can still set the wholesale price to influence the channel profit, but the appropriate value of \( r \) depends on \( \theta \), so perfect coordination is not achievable when realization of \( \theta \) is uncertain.

Previous studies have shown that perfect coordination can be derailed either because the vendor’s interest differs from that of the channel (i.e., double marginalization (Spengler (1950))), or because retailers differ in their product feature, cost structure, or market share (Iyer (1998) and Ingene and Parry (2000)). In our model, the vendor is interested only in maximizing the channel profit and no asymmetry exists among retailers. Still, perfect coordination is not obtainable and as (8) shows, the profit gap strictly increases with the number of retailers \( (N) \) and the degree of demand uncertainty \( (\sigma(\theta)) \).

The proposition is derived under two conditions. The condition of non-decreasing hazard rate \( f(\theta)/\bar{F}(\theta) \) is satisfied by most commonly-used distribution functions (uniform, exponential, and normal). The other condition, \( 2N\theta \geq (N-1)\mathbb{E}[\theta] + (N + 1)C \), is introduced to simplify the expression of the profit gap. Its removal would merely complicate the formula without changing the substance of our conclusion.
Having identified the profit gap, we now explain how the difference can be reduced by adopting the capacitated design. Following the description in Introduction, the design builds all hardware components at the beginning and activates them later as needed. Therefore, the design can be characterized by three parameters, $K$, the maximum system capacity, $c_f$, the per-unit hardware cost of building the maximum capacity in the first period, and $c_s$, the cost of activating installed capacity in the second period. Observe that $c_s$ is analogous to $C$ in the open-system design and $c_f K$ should be viewed as the amount of the first-period investment in excess of the fixed cost of the open-system design (which has been normalized to 0).

One might expect that by deploying all hardware at once, the capacitated system is cheaper to build. As we mentioned in Introduction, this economy of scale is the ultimate reason for recommending such design in cost-minimization models. However, to highlight our argument, we will ignore this benefit by assuming

$$c_f + c_s = C.$$  \hfill (9)

Therefore, the capacitated design differs from the open-system design in that 1) it shifts a part of capacity investment from the second period (after realization of $\theta$ is observed) to the first period (when $\theta$ is unknown), and 2) it imposes a maximum capacity limit. Both changes sound unhelpful, but as we explain next, the second effect is actually advantageous to maximizing channel profit, and if the advantage is strong enough to offset the negative impact of the first change, the capacitated design will become the desirable choice.

Under the capacitated design, the intra-channel competition model is

$$\max_{q_i} \left\{ \left[ \theta - \sum_{i=1}^{N} (q_i \wedge K) - r \right] (q_i \wedge K) \right\}, \quad i = 1, ..., N;$$  \hfill (10)

where $\wedge$ is the minimum operand, so $(q_i \wedge K)$ means retailer $i$’s output $q_i$ ($i = 1, ..., N$) does not exceed the maximum capacity $K$. Analogous to (3) and (4) in the open-system design case, the equilibrium output (per retailer) and retail price are

$$q^c(r, K, \theta) = \left( \frac{\theta - r}{N + 1} \wedge K \right)^+.$$
\[ p^c(r, K, \theta) = \theta - N \left( \frac{\theta - r}{N+1} \land K \right)^+ \] 

(11)

respectively; analogous to (5), the equilibrium profit (per retailer) is

\[ \pi^c(r, K, \theta) = \left[ p^c(r, K, \theta) - r \right] q^c(r, K, \theta); \]

and analogous to (6), the expected channel profit is

\[ \Pi_c(r, K) = \mathbb{E}[\pi^c(r, K, \theta) + (r - c^c)q^c(r, K, \theta)] - c^1 K \]

\[ = \mathbb{E} \left\{ \left( \theta - N \left( \frac{\theta - r}{N+1} \land K \right)^+ - c^c \right) \left( \frac{\theta - r}{N+1} \land K \right)^+ \right\} - c^1 K, \] 

(12)

which the vendor maximizes by choosing both \( r \) and \( K \), i.e.,

\[ \Pi^*_c = \max_{r, K} \{ \Pi_c(r, K) \}. \] 

(13)

**Proposition 2.** Suppose \( r^o \) optimizes (7), then there exists \((r^c, K^*)\) that maximizes (13), where

\[ r^c < r^o \text{ and } K^* < \frac{\tilde{\theta} - r^c}{N+1}. \] 

(14)

The proposition concludes that under the capacitated design, it is optimal to 1) charge retailers a strictly lower wholesale price than the price under the open-system design; and 2) allow retailers to be tightly constrained by the capacity limit in some cases (i.e., when \( \theta \) is close to \( \tilde{\theta} \)). It is tempting to explain both results by the cost difference between the two designs. After all, it is intuitive that when the per-unit cost of expanding capacity in the second period is reduced from \( C \) to \( c^* \), the wholesale price should fall as well. Moreover, when the capacity investment \( c^1 K \) is made in the first period before \( \theta \) is known, it is also sensible not to determine \( K \) based on the best case scenario, so the capacity limit should be binding in some cases. However, this explanation runs into a difficulty when we notice that both inequalities in (14) are strict and apply to all situations, including the (unlikely) case when \( c^1 = 0 \) (so according to (9), \( c^* = C \)). In other words, even if both designs have identical costs, their wholesale prices still differ; and even if it costs nothing to set the capacity limit to infinity under the capacitated design, it is still optimal to have capacity shortage.
The observation suggests that what (14) captures is not a mere consequence of cost difference, but a strategic exploitation of features related to the capacitated design.

To explain this strategy, compare the two competitive equilibrium in (3), (4), and (11). When $r^o > r^c$,

$$p^c(r^c, K, \theta) < p^o(r^o, \theta) \quad \text{and} \quad q^c(r^c, K, \theta) > q^o(r^o, \theta)$$

for all $\theta < (N + 1)K^* - r^c$, i.e., when demand is too weak to reach the capacity limit, the capacitated design results in a lower retail price and more sales to end users, which suits the situation well. The same outcome can be replicated under the open-system design if the vendor reduces the wholesale price from $r^o$ to $r^c$, but doing would also induce a low retail price if demand turns out to be strong and a higher retail price is needed to gain more profit. This dilemma is avoided under the capacitated design, which sustains a higher retail price in a strong-demand environment by keeping the total supply of capacity below a given limit.

One may perceive some similarities between the capacitated design and a direct resale price maintenance scheme in terms of their intended impact. In the latter case, the vendor explicitly specifies a minimum retail price in its contract with retailers (Katz (1989)). Both approaches can keep the retail price high when demand is strong, but the capacitated design does more. With high channel profit in high-demand scenarios secured by the capacity limit, the vendor can afford to lower the wholesale price, which induces a lower retail price and more profit in a weak demand environment. In contrast, the direct price maintenance scheme may simply collapse in this situation if no end user is willing to pay the minimum price.

The cost of pursuing the capacitated design is the inevitable waste associated with building the system to its maximum capacity up-front. There will always be cases when demand is weak enough to cause the equilibrium output to fall strictly below the limit, so some installed hardware will never be used. As is shown in the following proposition, the capacitated design is a better choice if the up-front investment does not reach some threshold.
Proposition 3. Let $\Pi^*_o$ and $\Pi^*_c$ be the optimal expected channel profits under the open-system and capacitated designs respectively. Suppose the cost equivalence condition (9) applies. Then there exists a strictly positive upper bound, $\bar{\epsilon} > 0$ such that

$$\Pi^*_o \leq \Pi^*_c \text{ if and only if } c_f \leq \bar{\epsilon}. \quad (15)$$

To summarize, the capacitated design gives rise to two types of inflexibilities: the inflexibility to expand capacity beyond a preset limit, and the inflexibility of requiring irreversible capacity investment ($c_f K$) before actual demand is known. Our discussion above shows that while investment inflexibility is detrimental to the channel profit, capacity inflexibility is actually beneficial. Proposition 3 tells us that which design is better depends on which type of inflexibility dominates. This trade-off is very different from the one considered in previous studies, where both inflexibilities are viewed negatively and can be tolerated only if they are associated with sufficient economy of scale.

4. Other Issues

While the primary purpose of channel coordination is to maximize profit, the consequence on the society and consumers also needs to be considered (see Section 5 of Katz (1989) for a survey). In our case, we need to decide which design delivers more benefit to the society and consumers. The following numerical example shows that the question may not have a definitive answer and the conclusion can be swayed by preference.

As in the last section, for the open-system design, let $r^o$ be the wholesale price, $q^o(\theta)$ be the output per retailer, and

$$p^o(\theta) = \theta - N q^o(\theta)$$

be the retail price. By definition, social welfare is

$$SW^o(\theta) = \int_0^{N q^o(\theta)} (p(q) - C) dq = \int_0^{N q^o(\theta)} (\theta - q - C) dq,$$

and consumer surplus is

$$CS^o(\theta) = \int_0^{N q^o(\theta)} (p(q) - p^o(\theta)) dq = \int_0^{N q^o(\theta)} (N q^o(\theta) - q) dq.$$
For the capacitated design, denote the optimal capacity level by $K^*$, wholesale price by $r^c$, output per retailer by $q^c(\theta)$, and the retail price by $p^c(\theta)$, so social welfare is

$$SW^c(\theta) = \int_0^{Nq^c(\theta)} (p(q) - c^s)dq - Nc^fK^* = \int_0^{Nq^c(\theta)} (\theta - q - c^s)dq - Nc^fK^*,$$

and consumer surplus is

$$CS^c(\theta) = \int_0^{Nq^c(\theta)} (p(q) - p^c(\theta))dq = \int_0^{Nq^c(\theta)} (Nq^c(\theta) - q)dq.$$

In the example, the number of retailers $N = 10$ and $\theta$ is uniformly distributed over $[\underline{\theta}, \bar{\theta}]$, where $\underline{\theta} = 5.6$ and $\bar{\theta} = 14.4$, and

$$C = 2, \ c^f = 0.7, \ c^s = 1.3.$$

Figure 1(a) compares social welfare for each realization of market condition $\theta$. We observe a threshold that separates low $\theta$ values at which $SW^c(\theta)$ dominates $SW^o(\theta)$ and high $\theta$ values at which the reverse is true. The same observation can be made about consumer surplus, which is compared in Figure 1(b). Both results make intuitive sense. Because the retail price is lower and more end users are served in under the capacitated design when demand is weak, social welfare and consumer surplus should improve. On the other hand, because the design restricts the supply of capacity to keep a high retail price when demand is strong both social welfare and consumer surplus are reduced in these cases. As for regulators and consumer protection groups, which design they should prefer really depends on what they think is more important. Given that demand is often associated with the general economic conditions, the capacitated design is a concern when the policy focus is on avoiding capacity bottleneck during economic booms. It becomes the favored choice if the main interest is to stimulate the industry and guarantee affordable service to consumers during tough economic times.

In some cases, channel management involves not only incentive alignment (for profit maximizing) and transfer payment (through the fixed fee in the two-part tariff), but also an “insurance” that guarantee a minimum profit to risk-averse retailers in the presence of market uncertainty (Rey and Tirole (1986)). Using the same numerical example as above, we find the capacitated design has
an “insurance” effect in reverse as it transfers profit volatility from the vendor to retailers. As is shown in Figures 2, the vendor’s profit becomes fixed and is below the profit under the open-system design when demand is sufficiently strong ($\theta$ is high) and is higher than the open-system design when demand is weak ($\theta$ is low). On the other hand, for retailers, both the upside reward (when $\theta$ is high) and downside risk (when $\theta$ is low) are increased.

The observed effect is consistent with the implication of the capacitated design as described in Proposition 2. In a strong market, sales are capped at $K^*$, so the vendor gets no more profit and retailers keep all gains derived from a higher retail price. On the other hand, as capacity is sold at a lower retail price ($r^*$), the vendor needs a higher fixed fee in the two-part tariff to collect profit.
and is able to do so under the perfect bargaining power. Consequently, it is the vendor who is more isolated from profit shortfalls in a weak market and retailers who absorb a larger share of risk. One may argue such outcome is desirable because having already taken R&D risks of developing and building a new system, the vendor should get a safer return by leaving profit volatility to retailers. As a counter argument, one may also contend that a vendor in monopoly is more capable of taking more risk than retailers in competition.

5. Conclusion

We discuss a strategic use of the capacitated system design to improve channel coordination. Even though our approach imposes vertical restrictions only implicitly by specifying the maximum system capacity, it has some advantages over direct mechanisms that dictate vertical restraints by contract terms. For instance, the design not only induces a high retail price when demand is strong but also keeps the price low when demand is weak, and thus improves the channel profit in both cases. In comparison, direct resale price maintenance is only suitable in the first situation. The design also mitigates the intra-channel competition by restricting the maximum output per retailer. A direct mechanism may achieve a similar outcome by explicitly dividing territories among retailers (Rey and Tirole (1986)), which causes a loss of positive externality and economy of scale that retailers can derive from operating in a larger market. The profit improvement is achieved by ex ante capacity commitment rather than ex post fixing (the latter gives rise to the question of legality), and thus allows social welfare and consumer surplus to improve in some cases to compensate for their declines in other cases. As its shortcoming, the design requires making an irreversible capacity investment in the presence of demand uncertainty, so it should be adopted if and only if the amount of inflexible investment does not exceed some threshold as given in Proposition 3.

In practice, technology progress has provided limitless possibility to expand network capacity, which spurred a tremendous growth of the communication services. However, many firms that build the networking infrastructure or deliver communication capacity are in financial trouble.
Part of the reason is because the technical power to expand has not been properly tamed by a suitable business model, resulting in firms be driven into a profit-losing game of capacity oversupply, which is summarized succinctly by the phrase “bandwidth glut” (Holzemer (2002)). Our analysis provides helpful insight on this problem by showing how channel profit can be improved by designing network systems intelligently. Moreover, as we show in Section 4, the gain for firms does not always represents a loss to consumers. In weak demand environment, both social welfare and consumer surplus improve

This approach is not obvious. The vendor may slip into a Bertrand Supertrap if it delegates the design decision to a R&D manager who is only responsible for technology excellence and cost minimization. The pitfall can be avoided only if the vendor takes a cross-function approach to understand broader implications of the design decision on market dynamics.

Our model is based on the assumption of a monopoly vendor, who might be a technology leader in the industry. An important extension is to consider the case when there are competing vendors. Our initial investigation shows that, in this case, the capacitated design may be over used as a tool to uncut competitors in an inter-channel competition. However, the outcome is driven by market dynamics orthogonal to that of the monopoly vendor case here, so we leave it to future work.

References


6. Appendix: Proof of Propositions

Proof for Proposition 1

The expected channel profit ultimately comes from sales to end users. When the chain is controlled perfectly by an integrated monopolist, then

$$\Pi^* = \mathbb{E} \left[ \max_q \{(\theta - q - C)q\} \right] = \frac{\mathbb{E}[(\theta - C)^2]}{4} = \frac{(\mathbb{E}[\theta] - C)^2 + \sigma^2(\theta)}{4}. \quad (16)$$

When the chain is separately controlled by an upstream vendor and downstream retailers, (3) and (4) show the expected profit,

$$\Pi_o(r) = \mathbb{E}[(p^o(r, \theta) - C)q^o(r, \theta)] = \mathbb{E} \left[ \left( - \frac{N(\theta - r)}{N+1} - C \right) \left( \frac{\theta - r}{N+1} \right)^+ \right].$$

We first consider the case $r \leq \theta$, in which case $\Pi_o(r)$ in the above can be expanded into

$$\Pi_o(r) = \frac{N}{(N+1)^2} \left[ -r^2 + \frac{(N-1)\mathbb{E}[\theta] + (N+1)C}{N} r + \frac{\mathbb{E}[\theta(\theta - (N+1)C)]}{N} \right]. \quad (17)$$
Maximize the above by controlling \( r \), the optimal solution is

\[
r^o = \frac{(N - 1)E[\theta] + (N + 1)C}{2N}.
\] (18)

Use \( r = r^o \) in (17) and collect terms, the optimal profit

\[
\Pi^*_o = \frac{N^2}{(N + 1)^2} \left[ \left( \frac{(N - 1)E[\theta] + (N + 1)C}{2N} \right)^2 + \frac{E[\theta - (N + 1)C]}{N} \right] + \frac{(E[\theta] - C)^2}{4} + \frac{N}{(N + 1)^2} \sigma^2(\theta) \geq 0.
\]

Compare the above with (16),

\[
\Pi^{**} - \Pi^*_o = \frac{(N - 1)^2}{4(N + 1)} \sigma^2(\theta).
\]

To complete the proof, we also need to exclude the case where \( r > \theta \), which is carried out by showing \( \Pi_o(\theta) \geq \Pi_o(r) \) for all \( r > \theta \). In the latter case, \( \theta \) varies from \( r \) to \( \bar{\theta} \), so in (7),

\[
\Pi_o(r) = \int_r^{\bar{\theta}} \left[ \left( \frac{\theta + Nr}{N + 1} - C \right) \left( \frac{\theta - r}{N + 1} \right) \right] f(\theta)d\theta,
\]

which differs from \( \Pi_o(\theta) \) by an amount of

\[
\Pi_o(\theta) - \Pi_o(r) = \int_r^{\bar{\theta}} \left[ \left( \frac{\theta + N\theta}{N + 1} - C \right) \left( \frac{\theta - r}{N + 1} \right) \right] f(\theta)d\theta - \int_r^{\bar{\theta}} \left[ \left( \frac{\theta + Nr}{N + 1} - C \right) \left( \frac{\theta - r}{N + 1} \right) \right] f(\theta)d\theta
\]

\[
= \int_r^{\bar{\theta}} \left( \frac{\theta + N\theta}{N + 1} - C \right) \left( \frac{\theta - r}{N + 1} \right) f(\theta)d\theta + \int_{\theta}^{r} \left( \frac{\theta - r}{N + 1} \right) f(\theta)d\theta + \int_{\theta}^{\bar{\theta}} \left( \frac{r - \theta}{N + 1} \right) f(\theta)d\theta
\]

Under the assumption \( \theta \geq C \), the first term is always positive. To show the second term is also positive, we apply the assumption

\[
2N\theta \geq (N - 1)E[\theta] + (N + 1)C
\] (19)

to replace \( C \), so

\[
\int_r^{\bar{\theta}} \left( \theta - N\frac{2\theta - r - \theta}{N + 1} - C \right) f(\theta)d\theta \geq \frac{1}{N + 1} \int_{r}^{\bar{\theta}} \left( -(N - 1)(\theta - E[\theta]) + N(r - \theta) \right) f(\theta)d\theta,
\]

and we can complete our proof by showing that the integral on the right-hand side

\[
\Lambda(r) = \int_r^{\bar{\theta}} \left( -(N - 1)(\theta - E[\theta]) + N(r - \theta) \right) f(\theta)d\theta
\]
is positive. Notice that $\Lambda(\theta) = \Lambda(\bar{\theta}) = 0$, so $\Lambda(r) \geq 0$ if it is either increasing or unimodal (first increasing, then decreasing in $r$). This is true because

$$
\frac{d\Lambda(r)}{dr} = NF(r) - [(r - E[\theta]) + N(E[\theta] - \bar{\theta})]f(r) = \bar{F}(r) \left[ N - [(r - E[\theta]) + N(E[\theta] - \bar{\theta})] \frac{f(r)}{F(r)} \right].
$$

Since $(r - E[\theta]) + N(E[\theta] - \bar{\theta})$ and $f(r)/\bar{F}(r)$ both increase in $r$, $d\Lambda(r)/dr$ cannot become positive again after its value reaches 0.

**Proof for Proposition 2**

We first show that to optimize (13), it suffices to consider

$$
\frac{\theta - r}{N + 1} \leq K \leq \frac{\bar{\theta} - r}{N + 1}.
$$

If $K > (\bar{\theta} - r)/(N+1)$, in (13), $(\theta - r)/(N + 1) < K$ for all $\theta$, so

$$
\Pi_c(r, K) = \mathbb{E} \left[ (\theta - NK - c^s)K \right] - c^f K,
$$

which can always be improved by reducing $K$. If $K < (\bar{\theta} - r)/(N + 1)$, in (13), $(\theta - r)/(N + 1) > K$ for all $\theta$, so

$$
\Pi_c(r, K) = \mathbb{E} \left[ (\theta - NK - c^s)K \right] - c^f K,
$$

which gives the same value as the case when $K = (\bar{\theta} - r)/(N + 1)$. Given (20) holds, the rest of the proof is based on a new function defined as

$$
\varphi(r, k, \bar{\theta}) \equiv \mathbb{E} \left[ \left( \theta - N \left( \frac{\theta - r}{N + 1} \wedge k \right) \right)^+ \left( \frac{\theta - r}{N + 1} \wedge k \right)^+ \right] - (C - \bar{\theta})k
$$

where $r \geq 0$, $\bar{\theta} \geq 0$, and $(\theta - r)/(N + 1) \leq k \leq (\bar{\theta} - r)/(N + 1)$. It is easy to verify that

$$
\Pi_o(r) = \varphi \left( r, \frac{\bar{\theta}}{N + 1}, C \right) \quad \text{and} \quad \Pi_c(r, K) = \varphi (r, K, c^s),
$$

so the difference of the two profits can be expressed as

$$
\Pi_o(r) - \Pi_c(r, K) = \varphi \left( r, \frac{\bar{\theta}}{N + 1}, C \right) - \varphi (r, K, c^s)
= \int_{K}^{\theta/(N + 1)} \partial \varphi(r, k, c^s) \partial k \, dk + \int_{c^s}^{C} \partial \varphi (r, \theta/(N + 1), \bar{\theta}) \partial \bar{\theta} \, d\bar{\theta}. \quad (22)
$$
If we can show that for all \( r \leq r^c, k \geq K^* \) and \( \tilde{c} \geq c^s \),

\[
\frac{\partial \varphi(r, k, c^s)}{\partial k} > 0 \quad (23)
\]

and

\[
\frac{\partial \varphi(r, \tilde{\theta}/(N + 1), \tilde{c})}{\partial \tilde{c}} > 0, \quad (24)
\]

then \( r^o \geq r^c \) can be proved by contraction: if \( r^o < r^c \), then (23) implies that

\[
\frac{\partial \varphi(r^o, k, c^s)}{\partial k} < \frac{\partial \varphi(r^c, k, c^s)}{\partial k} \quad \text{for } k \geq K^*,
\]

and (24) implies that

\[
\frac{\partial \varphi(r^o, \tilde{\theta}/(N + 1), \tilde{c})}{\partial \tilde{c}} < \frac{\partial \varphi(r^c, \tilde{\theta}/(N + 1), \tilde{c})}{\partial \tilde{c}} \quad \text{for } \tilde{c} \geq c^s.
\]

Apply both to (22),

\[
\Pi_o(r^o) - \Pi_c(r^o, K^*) < \Pi_o(r^c) - \Pi_c(r^c, K^*).
\]

By rearranging terms,

\[
\Pi_o(r^o) + \Pi_c(r^c, K^*) < \Pi_o(r^c) + \Pi_c(r^o, K^*),
\]

which gives rise to a contradiction because \( \Pi_o(r) \) and \( \Pi_c(r, K) \) are supposed to be optimized by \( r^o \) and \( (r^c, K^*) \), respectively. To prove (23) and (24), the latter is almost immediate. By (22),

\[
\frac{\partial \varphi}{\partial \tilde{c}} = -E \left[ \left( \frac{\theta - r}{N + 1} \land k \right)^+ \right] + k,
\]

which increases in \( r \). Proving (23) is more involved. We first expand (21) into

\[
\varphi(r, k, \tilde{c}) = \int_{g_{r\theta}}^{(N+1)k+r} \frac{[\theta + Nr - (N + 1)\tilde{c}](\theta - r)}{(N + 1)^2} f(\theta) d\theta
\]

\[
+ \int_{(N+1)k+r}^{\theta} (\theta - Nk - \tilde{c})k f(\theta) d\theta - (C - \tilde{c})k. \quad (25)
\]

The related derivatives are

\[
\frac{\partial \varphi(r, k, \tilde{c})}{\partial r} = \int_{g_{r\theta}}^{(N+1)k+r} \frac{(N - 1)\theta - 2Nr + (N + 1)\tilde{c}}{(N + 1)^2} f(\theta) d\theta, \quad (26)
\]

and

\[
\frac{\partial \varphi(r, k, \tilde{c})}{\partial k} = [(N - 1)k - r + \tilde{c}] f((N + 1)k + r),
\]
so (23) holds if
\[(N - 1)K^* - r^e + c^e \geq 0. \tag{27}\]

To show this is true, because \(r^e\) optimizes \(\Pi_\circ(r, K^*) = \varphi(r, K^*, c^e)\),
\[
\frac{\partial \varphi(r, K^*, c^e)}{\partial r} = 0 \text{ at } r = r^e,
\]
so the integrant in (26) must be non-negative when \(k = K^*, r = r^e, \tilde{c} = c^*\) and \(\theta\) takes the maximum value of \((N + 1)K^* + r^e\), i.e.,
\[0 \leq -2N r^e + (N - 1)[(N + 1)K^* + r^e] + (N + 1)c^e = (N + 1)[(N - 1)K^* - r^e + c^e],\]
which completes our proof of (23), and thus that of \(r^e \leq r^o\). Finally, we prove \(K^* < (\bar{\theta} - r^e)/(N + 1)\).

Following (25), at \(k = (\bar{\theta} - r)/(N + 1)\),
\[
\frac{\partial \rho(r, k, \bar{c})}{\partial k} = -(C - \bar{c}) \quad \text{and} \quad \frac{\partial^2 \rho(r, k, \bar{c})}{\partial k^2} = (N + 1)[(N - 1)k - r + \bar{c}]f(\bar{\theta}).
\]

If \(K^* = (\bar{\theta} - r^e)/(N + 1)\), then \(\Pi_\circ(r^e, K^*) = \varphi(r^e, (\bar{\theta} - r^e)/(N + 1), c^e)\), so
\[
\frac{\partial \Pi_\circ(r^e, K^*)}{\partial K} = -(C - c^e) \quad \text{and} \quad \frac{\partial^2 \Pi_\circ(r^e, K^*)}{\partial K^2} = (N + 1)[(N - 1)K^* - r^e + c^e]f(\bar{\theta})
\]
Because \(C \geq c^e\) and because of (27),
\[
\frac{\partial \Pi_\circ(r^e, K^*)}{\partial K} \leq 0 \quad \text{and} \quad \frac{\partial^2 \Pi_\circ(r^e, K^*)}{\partial K^2} \geq 0,
\]
which violates 1st or 2nd order optimality conditions, optimal capacity \(K^*\) cannot be reached at \((\bar{\theta} - r^e)/(N + 1)\).

**Proof for Proposition 3**

Since \(\Pi^*_\circ\) does not change with \(c^f\), we prove the proposition by showing a) \(\Pi^*_\circ\) decreases in \(c^f\) and b) \(\Pi^*_\circ < \Pi^*_\circ\) if \(c^f = 0\). To prove a), if \(c^f_1 > c^f_2\), for any given \(r^e, K\), then (note \(c^e = C - c^f\) by (9))
\[
\Pi_\circ(r^e, K|c^f_1) - \Pi_\circ(r^e, K|c^f_2) = (c^e_2 - c^e_1)E\left[\left(\frac{\theta - r^e}{N + 1} \wedge K\right)^+\right] - (c^e_1 - c^f_1)K
\]
\[
= (c^e_1 - c^e_2)E\left[\left(\frac{\theta - r^e}{N + 1} \wedge K\right)^+\right] - K < 0.
\]

To prove b), \(c^e = C\) when \(c^f = 0\), so according to (7) and (13), \(\Pi^*_\circ = \Pi^*_\circ(r^e, \frac{\theta - r^e}{N + 1})\). From Proposition 2 we know \(K = \frac{\theta - r^e}{N + 1}\) is not optimal for \(\Pi_\circ(r^e, K)\), so \(\Pi^*_\circ < \Pi^*_\circ\).