

Incentives for Quality through Endogenous Routing

Lauren Xiaoyuan Lu, Jan A. Van Mieghem,
R. Canan Savaskan

July 14, 2006
COSM-06-002

Working Paper Series

Center for Operations and Supply Chain Management



Northwestern University

Incentives for Quality through Endogenous Routing

Lauren Xiaoyuan Lu · Jan A. Van Mieghem · R. Canan Savaskan

Kellogg School of Management, Northwestern University

July 14, 2006

Abstract

We consider quality control and rework routing policies of a firm implementing piece rate compensation. When a worker generates a defect, rework is conventionally assigned to the originating worker (in a *self routing* scheme) or to another worker dedicated to rework (in a *dedicated routing* scheme). In contrast, a novel *cross routing* scheme allocates any worker's defects to a parallel worker performing both new jobs and rework. All the workers receive the piece rate paid per job upon passing quality inspection or at rework completion. We compare the incentives of these different rework routing schemes by embedding quality control and routing of a multi-class queuing network in a principal-agent model. We show that conventional self routing of rework can never induce first-best effort. Dedicated and cross routing, however, can lead to higher profits for the principal and improve incentives for quality by imposing an implicit punishment for quality failure. In addition, cross routing leads to workload allocation externalities and a prisoner's dilemma between the two parallel workers, thereby creating the highest incentives for quality. In general, cross routing generates the highest profit rate when appraisal, internal failure, or external failure costs are high, while self routing performs the best when gross margins or disutilities of effort are high.

Key words: queuing networks; routing; Nash equilibrium; quality control; piece rate.

1 Introduction

This paper investigates how incentives and endogenous rework routing can induce effort and improve a firm's quality and profits. It is motivated by the practice of a service operations firm, Memphis Auto Auction, who is a wholesale automotive liquidator of used vehicles and employs two teams of workers that clean and detail vehicles in parallel. The workers are paid piece rate while the quality control leader is paid salary plus a bonus based on overall work quality. The firm ties worker pay to quality through an unconventional rework routing scheme illustrated in Figure 1C that we will call *cross routing*. If the vehicle passes quality inspection, the team earns the piece rate. Otherwise, the

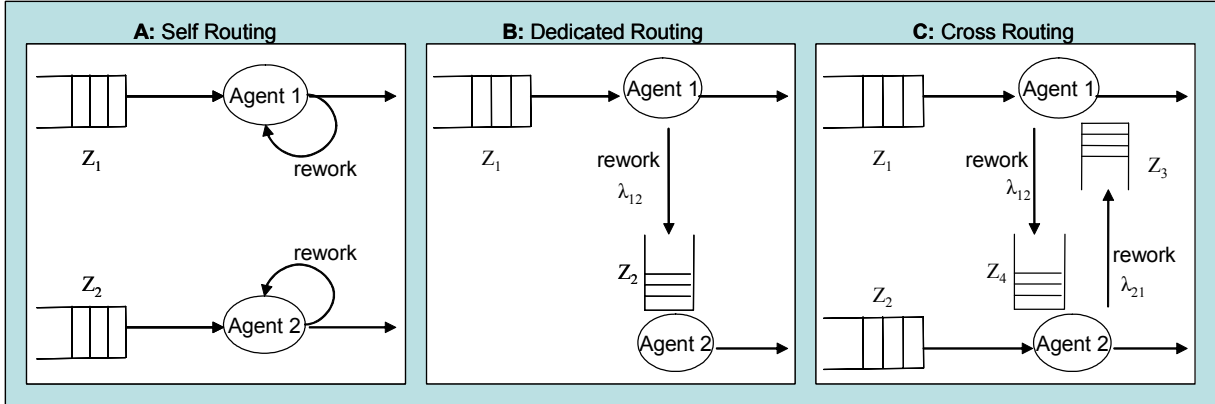


Figure 1: Three Rework Routing Schemes: (A) Self Routing, (B) Dedicated Routing, and (C) Cross Routing.

vehicle’s rework is routed to the other team who earns the piece rate at rework completion (while the first team gets paid nothing). This cross routing of rework contrasts with the conventional practices that assign rework back to the originating team (Figure 1A) or to a dedicated rework team (Figure 1B). In the *self routing* scheme, each team has to conduct rework without additional pay (i.e., only a single piece rate is paid per job). In the *dedicated routing* scheme, however, the originating team loses the pay on the defective job completely. The two conventional rework routing schemes are natural choices often used in practice. We shall show that these three schemes differ not only in rework routing policies but also in their induced first-pass quality levels.

Our main research goal is to develop an in-depth understanding of how these three routing and incentive schemes compare in terms of quality and firm profits. We address this question by embedding quality control and routing of a multi-class queuing network in a principal-agent model. Rework routing impacts agent incentives to exert quality-improving effort in two important ways. First, self routing gives the agents a second chance to work on a job and earn the piece rate, resulting in a disincentive for the agents to exert first-pass effort. In contrast, dedicated and cross routing implicitly punish the agents for quality failure by allocating rework to a different agent, thereby boosting the incentives for first-pass effort.

Second, whereas self routing gives each agent independent and direct control over the workload of new jobs and rework, the agents’ workload in cross routing is determined by the equilibrium outcome of the effort game played between them. When rework takes less effort than first-pass work, agents prefer rework. To raise their utilization on rework, the agents in cross routing increase their first-pass effort as a result of the workload allocation externality arising in the effort game.

We explain this subtle effect by characterizing the strategic interaction between the agents and the dynamics of the queuing network. In a capacity constrained system, both agents are continuously busy working either on new jobs or rework. One agent is able to direct the other agent's work allocation toward new jobs by routing less rework to him. Keeping effort unchanged, the parallel agent generates more rework in return. Since his pay rate suffers from low rework inflow, the parallel agent counteracts by increasing his first-pass effort. Consequently both agents exert high first-pass effort and receive low rework allocation in equilibrium. This equilibrium exhibits a prisoner's dilemma, where each agent has an incentive to exert high effort when the other agent exerts low effort, even though both agents would jointly benefit from an cooperative outcome of both exerting low effort. Nevertheless, the agents are compensated with higher piece rates for their increased effort.

Though higher first-pass effort produces fewer internal and external defects, it does not always lead to higher profits for the principal. Inducing first-pass effort benefits the principal by improving quality and reducing three of the four quality costs in Juran's cost-of-quality framework (Juran & Gryna (1993)): internal failure costs, external failure costs, and appraisal costs. On the other hand, higher first-pass effort implies higher piece rate compensation and lower throughput of the system as the agents spend more effort (processing time) per job in expectation. Since piece rate compensation cost can be deemed as a form of prevention costs, our model covers all of the four dimensions of the cost-of-quality framework. It predicts that the principal would strive for the optimal defect rate (which has a one-to-one relationship to the induced first-pass effort) to achieve the lowest cost by balancing the costs of non-conformance with appraisal and prevention costs. Built on this cost minimization view of quality management, our model adds an additional dimension: throughput and thus revenues also impact a firm's quality control decisions. Since high effort leads to low throughput in a capacitated system, the principal must trade-off quality with revenue.

The endogeneity of rework routing in the principal's decision model has two aspects. First, the principal compares the profit rate of the three routing schemes and implements the most favorable one. Second, once the principal chooses a routing scheme, the agents' actual workload balance between new jobs and rework is determined by their effort levels, which are induced by the principal's incentive and quality inspection decisions.

Using an analytical model we establish the following results. Conventional self routing of rework can never induce agents to exert first-best effort. Dedicated and cross routing of rework, however, offer some remedy by inducing higher effort and quality, which can lead to higher profits for the

principal. We show that at any quality inspection precision, dedicated and cross routing induce higher effort than self routing. As a result, piece rates paid in these two schemes are higher. Financial performance crucially depends on quality-related costs, gross margins, and agent disutilities of effort. In general, cross routing of rework generates the highest profit rate when the appraisal, internal failure, or external failure costs are high, i.e., when it is important to induce high first-pass effort. On the other hand, self routing of rework performs the best when the gross margin or agent disutility of effort are high, i.e., when throughput is important or labor is costly.

The remainder of the paper is organized as follows. Section 2 reviews related literature while Section 3 lays out the main model. Sections 4 and 5 analyze the networks with excess capacity and capacity constraint, respectively. In each of the two sections, we first derive the first-best benchmark and then analyze the three rework routing schemes, and finally compare their performance. In the rest of the paper, we will use superscripts S , D , and C to denote self, dedicated, and cross routing, respectively. In addition, we will use superscript FB to denote the first-best solutions.

2 Related Literature

This paper contributes to three streams of literature, each of which we briefly review below. The first stream of research relevant to our work is the economics literature on compensation and job design, which studies the moral hazard problem that arises when a worker's effort is imperfectly observed. Worker compensation is thus often based on output instead of effort. Holmstrom & Milgrom (1991) explain the trade-offs between inducing effort towards quantity vs. quality with a multitask principal-agent model. In their model, producing high volume and good quality output is viewed as two dimensional tasks of a worker's job. They argue that it would be costly, if not impossible, to achieve good quality with piece rate compensation if quality were poorly measured. Instead of taking a multitask approach, we manifest the intrinsic trade-off between quantity and quality with a single dimensional decision variable, i.e., the average processing time spent on each job. Moreover, we provide theoretical support that smart routing of rework is capable of inducing quality-improving effort even under piece rate compensation. Lazear (2000) provides empirical evidence that adoption of piece rate significantly improves productivity. In Lazear's real-world example, rework is assigned to the originating worker (i.e., self routing) and quality does not deteriorate after the firm implements piece rate compensation scheme. He argues that workers have incentives to get it right in the first time because rework is costly. In contrast, we will show

that workers always exert system suboptimal quality effort in the self routing scheme.

Holmstrom & Milgrom (1991) also demonstrate that job design is an important instrument for the control of incentives. They find that tasks should be grouped into jobs such that easily measured tasks are assigned to one worker and hard-to-measure tasks to the other worker. Though we use a one-dimensional principal-agent model, there are two tasks in our model that differ in their measurability: first-pass work is monitored imperfectly by quality inspection and rework has no uncertainty in output. Supporting Holmstrom & Milgrom (1991)'s theory that tasks should be separated according to their measurability characteristics, the dedicated routing scheme achieves advantageous incentive power over the self routing scheme.

The second relevant stream of literature is on the economics of quality control and inspection in a game-theoretic setting. Note that these papers consider quality-related contractual issues between firms and are only tangentially related to our work. For example, Reynier & Tapiero (1995) study the effect of contract parameters and warranty costs on the choice of quality by a supplier and the quality control policy by a buyer. Baiman, Fischer & Rajan (2000) focus on how contractibility of quality-related information impact the product quality and profits of a supplier and a buyer. Our work studies how rework routing and costs of quality affect the workers' choice of quality-improving effort and the firm's quality inspection policy.

Third, from a methodological perspective, we combine the two previous literatures on principal-agent models and quality management with that of queuing networks. Much of agency theory seeks contracts that maximize a principal's objective subject to an agent's post-contractual opportunistic behavior. However, little is known about quality control policies, i.e., how precisely should performance be measured? Queuing network models can capture system dynamics and quality inspection levels and allow us to draw operational insights that are largely missing in the existing agency literature. We endogenize quality control policy by allowing the principal to set quality inspection precision level. In addition, by considering capacity-constrained systems, we allow agents' effort levels (i.e., processing times) to directly impact system throughput, i.e., productivity. Similar work can be found in the literature that studies the impact of decentralized decision making on process performance in queuing systems. One of the seminal work dates back to Naor (1969), who studies how pricing can be used to achieve social optimum and prevent performance degradation as a result of customers' self-interested behavior. Other papers (e.g., Mendelson & Whang (1990), Ha (2001), Van Mieghem (2000), Afeche & Mendelson (2004), etc.) have also designed pricing mechanisms to achieve system optimal performance. None of these works model quality inspection and rework.

Principal-agent models in queuing systems have been explored in the operations management literature. A sample of recent papers include Gilbert & Weng (1998), Plambeck & Zenios (2000), Shumsky & Pinker (2003), and Gunes & Aksin (2004). Gilbert & Weng (1998) study the incentive effects of common vs. separate queue allocation schemes on self-interested operators' capacity decisions in a principal-agent model. Plambeck & Zenios (2000) study incentives in a dynamic setting where agents' effort influences the transition probabilities of a system. Similarly, in our model, probabilistic routing is determined by the agents' effort level. Our paper is closely related to Shumsky & Pinker (2003) in that the principal designs incentives to induce effort in steady state. Our work differs from Shumsky & Pinker (2003) in two important ways. First, we explicitly model the queuing network dynamics and also consider the case where the system is capacity constrained. Second, the principal in our model hires two agents whose expected utility rates are interdependent. Therefore, we need to investigate the strategic interactions between the agents and derive the effort Nash equilibrium. Gunes & Aksin (2004) model the interaction of market segmentation, incentives, and process performance of a service-delivery system using a single-server queue embedded in a principal-agent framework. Out of these papers, only Gilbert & Weng (1998) consider two servers, thus a network setting, which is the closest to our queuing network model. The novelty of our model in terms of incorporating a queuing system in a principal-agent framework lies in that we model two endogenous queues, i.e., the rework queues that are generated by the agents and the arrival rate of rework is endogenously determined by the agents' effort.

Examples of games in queuing systems can be found in Cachon & Harker (2002) and Parlakturk & Kumar (2004). Cachon & Harker (2002) investigate the competition dynamics between two service providers in a queuing game when outsourcing is allowed. Parlakturk & Kumar (2004) propose a scheduling rule that achieves first-best system performance in the presence of self-interested routing of customers. The queuing games in our model is distinct because the players are two agents whose effort directly impacts the capacity and quality output of the system, which in turn affects the principal's profit.

In our motivating example, Memphis Auto Auction employed teams to complete jobs. In this paper, we will treat teams as agents and ignore the team incentive issues that may arise due to free riding and collaboration. A relevant reference for team incentives is Hamilton, Nickerson & Owan (2003), which empirically investigates the impact of teams on productivity. They distinguish individual piece rate used in flow production from group piece rate used in modular production. They found that group piece rate has a stronger incentive effect on productivity than individual

piece rate due to collaboration among team members.

3 The Model

In this section we present the modeling constructs that drive the operational and the financial flows in our principal-agent model.

Operational Flows. Consider an operation where a principal hires two identical risk neutral agents to complete work (“jobs”) and subsequently inspects their output quality. The principal sets quality inspection precision $p \in [0, 1]$, which is the probability of catching a defect given a bad output (A good quality output passes inspection with probability 1). Each agent chooses first-pass effort (processing time) t , where $t \geq \underline{t}$ and $\underline{t} > 0$ is the minimum effort that can be exerted. We assume that the agent’s service time of each job is exponentially distributed with mean t . This strategic decision variable of the agent is not observable by the principal, but drives the quality of the output. Let $F(t)$ denote the probability of producing good quality given first-pass effort t , with $F(\underline{t}) = 0$ and $F(\infty) = 1$. We assume that F is strictly concave and increasing (i.e., $F'' < 0$, $F' > 0$), and denote $f = F'$ and $\bar{F} = 1 - F$. Upon identifying defects, the principal routes the rework either to the originating agent in self routing, to the agent dedicated to rework in dedicated routing, or to the parallel agent in cross routing. We assume that rework always generates good output, thus poor quality only results from not catching the first-pass defects. The overall quality conformance level that an external customer experiences is then calculated as $F(t) + p\bar{F}(t)$, which we will use to indicate the principal’s quality output, denoted by Q . Therefore, at any inspection precision $p < 1$, higher inspection precision or higher first-pass effort result in higher quality output.

We will show that the incentive effects of the three routing schemes crucially depend on whether the network is capacitated or not. With excess capacity, each agent is supplied with Poisson arrival of new jobs. Therefore, the agents have idle time and their effort levels do not impact throughput. In contrast, when the system is capacity constrained, agents are continuously busy and their effort levels directly impact throughput (productivity).

For tractability, we assume that rework takes r units of time on average, where r is common knowledge. Since defects have to be corrected as instructed by the principal, we assume rework effort is observable, i.e., no moral hazard problem in rework. We argue that even if agents may exhibit opportunistic behavior in performing rework, the effect is limited because identified defects have to be corrected completely. We assume that rework time is exponentially distributed with

mean r . Furthermore, rework has preemptive priority over new jobs. This priority rule is adopted because of two considerations. First, in a capacitated system, the agents can be always engaged in new jobs. Without the priority rule, defects may never be reworked. Second, priority rule simplifies analytics of the model. Finally, we assume that rework takes less time than the minimum first-pass processing time:

$$r \leq t. \tag{A1}$$

This assumption allows us to focus on the interesting range of parameter values that highlight the moral hazard problem and the efficacy of “smart” rework routing in inducing effort. We will discuss the implications of this assumption when we compare the performance of the three rework routing schemes in Section 4.5.

Financial Flows and Incentives. Each agent earns a piece rate w for a completed job that passes quality inspection or for a rework. The agents’ disutility of effort per unit time is α . Without loss of generality, we normalize the agents’ reservation utility to be 0. In a competitive labor market, α can also be interpreted as the outside wage rate. The principal earns gross margin v per completed job that passes quality inspection, pays agents, and incurs three quality costs classified as Juran’s cost-of-quality framework: (1) an appraisal cost per new job denoted by $C_A(p)$. We assume $C_A(0) = C'_A(0) = 0$ and $C'_A(1) = \infty$, which implies that in equilibrium the principal chooses an interior inspection policy, i.e. $p \in (0, 1)$. In addition, $C'_A > 0$, $C''_A > 0$. Note that these are standard assumptions frequently used in the quality management literature (e.g., Baiman et al. (2000)). (2) an expected internal failure cost per new job, denoted by $C_I(p, t) = p\bar{F}(t)c_I$, where c_I is the cost per defect when internally detected; (3) an expected external failure cost per new job, denoted by $C_E(p, t) = (1 - p)\bar{F}(t)c_E$, where c_E is the cost per defect when externally detected. (External failure costs are typically larger than internal failure costs: $c_E > c_I$. Otherwise, the principal would have no incentives to fix defects internally.) We assume that the principal maximizes her long-run average profit rate, denoted by V , while the agents maximize their long-run average utility rate, denoted by U .

The incentive contract offered by the principal consists of two elements: the quality inspection precision p and the piece rate w . It is worth noting that we intentionally restrict the principal’s choice of contract to piece rate. This modeling choice is motivated by the fact that piece rate has been empirically demonstrated effective in improving productivity. Moreover, consistent with the personnel economics literature, piece rate is often used when productivity is important and quality

monitoring is possible. Our objective is to evaluate real-world practices that involve both piece-rate compensation and quality inspection.

4 Case I: Excess Capacity

Excess capacity implies that the principal maintains a sufficient staffing level to complete all jobs and the agents have idle time in steady state. Hence, the throughput of the system is driven by the exogenous market demand, which is represented by the arrival rate of jobs (denoted by 2λ). The principal focuses on reducing internal and external failure costs through quality inspection and inducing first-pass effort, while controlling for appraisal costs and agent compensation costs. Let ρ_i denote the utilization of agent i . Throughout this section, we assume the system is stable in steady state. The stability condition for the system is $\max_{i \in \{1,2\}} \rho_i < 1$.

4.1 The First-Best Benchmark

When effort is observable, the principal's problem is independent of whether rework is performed by the originating agent or a different agent. For expositional convenience, we derive the first-best benchmark using the self routing scheme. The agents spend on average $t + p\bar{F}(t)r$ time units per job. Since the job arrival rate is λ per agent, renewal theory yields that the agents' long-run average utility rate is $\lambda[w - \alpha(t + p\bar{F}(t)r)]$. Though the principal hires two agents, the contracting problem of each agent is independent and identical. The principal maximizes her long-run average profit rate:

$$V^{FB} = \max_{0 \leq p \leq 1, w \geq 0, t \geq \bar{t}} 2\lambda[v - w - \bar{F}(t)(pc_I + (1-p)c_E)] - C_A(p), \quad (1)$$

$$\text{subject to} \quad \lambda[w - \alpha(t + p\bar{F}(t)r)] \geq 0 \quad (\text{IR}). \quad (2)$$

The individual rationality (IR) constraint specifies the agents' outside option. Note that we define only one IR constraint because it is the same for the two identical agents. Since the principal's profit rate is monotonically decreasing in the piece rate w , the IR constraint must bind, simplifying the principal's problem to an optimization problem of two variables: t and p . Let $\{t^{FB}, p^{FB}\}$ denote the first-best solution to the above optimization problem¹ with observable effort. Since $\rho_i = \lambda(t^{FB} + p\bar{F}(t^{FB})r)$, the stability condition becomes $\lambda < \frac{1}{t^{FB} + p\bar{F}(t^{FB})r}$. For a stable system,

¹We ignore the issue of uniqueness of solution as all of our subsequent results hold for any interior optimum.

Lemma 1 characterizes the first-best solution (proofs for all lemmas and propositions are presented in the Appendix).

Lemma 1 *If $c_E > c_I + \alpha r > \frac{\alpha}{f(\underline{t})}$, there exists an interior first-best solution $\{t^{FB}, p^{FB}\}$, which is characterized by*

$$f(t^{FB}) = \frac{1}{p^{FB}r + \frac{1}{\alpha}(p^{FB}c_I + (1 - p^{FB})c_E)}, \quad (3)$$

$$C'_A(p^{FB}) = \bar{F}(t^{FB})(c_E - c_I - \alpha r). \quad (4)$$

It is simple to show that $\frac{\partial^2 V}{\partial t \partial p} < 0$, i.e., t and p are strategic substitutes. Since the principal is the Stackelberg leader and the agent earns zero utility rate in equilibrium, the principal's objective is identical to a central planner's. Therefore, the first-best solution achieves the Pareto optimum for the entire system. Moreover, the resulting first-best piece rate² $w^{FB} = \alpha(t^{FB} + p^{FB}\bar{F}(t^{FB})r)$.

4.2 Self Routing

When effort t is not observable and the rework is routed back to the originating agent, the principal's problem becomes

$$V^S = \max_{0 \leq p \leq 1, w \geq 0} 2\lambda[v - w - \bar{F}(t)(pc_I + (1 - p)c_E)] - C_A(p), \quad (5)$$

$$\text{subject to } \lambda[w - \alpha(t + p\bar{F}(t)r)] \geq 0 \quad (\text{IR}), \quad (6)$$

$$t \in \arg \max_{t' \geq \underline{t}} \lambda[w - \alpha(t' + p\bar{F}(t')r)] \quad (\text{IC}). \quad (7)$$

The additional incentive compatibility (IC) constraint describes the agents' post-contractual optimization behavior. Since the two agents are completely independent and symmetric, we only need a single IR and IC constraint. Let t^S denote the agents' best response to the incentive contract $\{p, w\}$. The first-order condition is equivalent to

$$f(t^S) = \frac{1}{pr}, \quad (8)$$

which can be rewritten as $t^S(p) = f^{-1}(\frac{1}{pr})$. Hence, the stability condition becomes $\lambda < \frac{1}{t^S + p\bar{F}(t^S)r}$.

Lemma 2 *If $t^S(p) \geq \underline{t}$, then $t^S(p)$ is a unique global maximum and is the agents' best response to the incentive contract $\{p, w\}$.*

²Notice that we specify the first-best solution only in terms of t and p to represent system-optimal levels of effort and quality inspection, while w is only a transfer between the principal and the agents.

Since the agents have sufficient time to complete all jobs and always earn the piece rate of each job, the agents' optimal effort is not impacted by the job arrival rate λ and the piece rate w . However, the first-pass effort increases when the principal raises quality inspection precision level or when rework is costly to the agents.

4.3 Dedicated Routing

Under dedicated routing, rework is assigned to an agent dedicated to rework. Without loss of generality, we assign new jobs to agent 1 and rework to agent 2. To keep the system's supply of jobs unchanged, agent 1 is assigned with job arrival rate 2λ . The principal maximizes her long-run average profit rate:

$$V^D = \max_{0 \leq p \leq 1, w \geq 0} 2\lambda[v - w - \bar{F}(t)(pc_I + (1-p)c_E)] - C_A(p), \quad (9)$$

$$\text{subject to} \quad 2\lambda[(1-p)\bar{F}(t)w - \alpha t] \geq 0 \quad (\text{IR1}), \quad 2\lambda p\bar{F}(t)(w - \alpha r) \geq 0 \quad (\text{IR2}) \quad (10)$$

$$t \in \arg \max_{t' \geq \underline{t}} 2\lambda[(1-p)\bar{F}(t')w - \alpha t'] \quad (\text{IC1}). \quad (11)$$

Since agent 2's effort is fully observable as he only performs rework, only IC1 is needed. Let t^D denote agent 1's best response to the incentive contract $\{p, w\}$. Then, t^D satisfies

$$f(t^D) = \frac{\alpha}{pw}, \quad (12)$$

which can be rewritten as $t^D(p, w) = f^{-1}(\frac{\alpha}{pw})$. Agent 1 and 2's utilizations are $\rho_1 = 2\lambda t^D$ and $\rho_2 = 2\lambda p\bar{F}(t^D)r$, respectively. The stability condition becomes $\lambda < \min\{\frac{1}{2t^D}, \frac{1}{2p\bar{F}(t^D)r}\}$.

Lemma 3 *If $t^D(p, w) \geq \underline{t}$, $t^D(p, w)$ is a unique global maximum and is agent 1's best response to the incentive contract $\{p, w\}$.*

Now agent 1's best response depends on both p and w . Therefore, the principal can induce higher first-pass effort not only by increasing the quality inspection precision but also by raising the piece rate.

4.4 Cross Routing

When rework is assigned to the parallel agent, a rework queue is generated and its queue size depends on the first-pass effort level of the originating agent. We now must characterize the rework equilibrium queues as part of the principal-agent incentive problem. For the multi-class queuing network illustrated in Figure 1C, we define the following rates for agent $i, j \in \{1, 2\}$ and $i \neq j$:

- Agent i 's new job service rate $\mu_i^n = \frac{1}{t_i}$
- Agent i 's rework service rate $\mu_i^r = \frac{1}{r}$
- Agent i 's defect generation rate (or agent j 's rework arrival rate) $\lambda_{ij} = \frac{p\bar{F}(t_i)}{t_i}$

Let a four-dimensional vector (Z_1, Z_2, Z_3, Z_4) represent the state of the four queues of the system (two new job queues and two rework queues). The detailed balance equations are too complex to be solved analytically in closed form. Luckily however, we do not need the limiting distribution of every single state to compute the utility rate of the agents. We only need to know the aggregate probabilities of the agents being idle π_i^0 , working on new jobs π_i^n , and working on rework π_i^r . In steady state, the queuing network must satisfy

$$\begin{aligned}
\pi_i^0 + \pi_i^n + \pi_i^r &= 1 \quad (\text{Law of total probability}), \\
\lambda &= \mu_i^n \pi_i^n \quad (\text{Balance of agent } i\text{'s new job queue}), \\
\lambda_{ji} \pi_j^n &= \mu_i^r \pi_i^r \quad (\text{Balance of agent } i\text{'s rework queue}),
\end{aligned} \tag{13}$$

for $i, j \in \{1, 2\}$ and $i \neq j$. Solving the above equations yields

$$\pi_i^0 = 1 - \lambda t_i - \lambda p \bar{F}(t_j) r, \quad \pi_i^n = \lambda t_i, \quad \pi_i^r = \lambda p \bar{F}(t_j) r \tag{14}$$

Agent i 's long-run average utility rate

$$\begin{aligned}
U_i(t_i, t_j) &= \pi_i^n \times \frac{(1 - p \bar{F}(t_i))w - \alpha t_i}{t_i} + \pi_i^r \times \frac{w - \alpha r}{r} \\
&= \lambda [(1 - p \bar{F}(t_i))w - \alpha t_i] + \lambda p \bar{F}(t_j)(w - \alpha r).
\end{aligned} \tag{15}$$

Notice that the first term is agent i 's average reward rate from working on new jobs, while the second term is his average reward rate from completing rework generated by agent j . Let t_i^C denote agent i 's best response. From the first-order condition, it follows that $f(t_i^C) = \frac{\alpha}{pw}$. Because $f(t_i^C) = f(t_j^C)$, we drop the subscript from now on:

$$f(t^C) = \frac{\alpha}{pw}, \tag{16}$$

which can be rewritten as $t^C(p, w) = f^{-1}(\frac{\alpha}{pw})$. Since agent i 's utilization $\rho_i = \lambda(t_i + p \bar{F}(t_j)r)$, the stability condition becomes $\lambda < \frac{1}{t^C + p \bar{F}(t^C)r}$.

Lemma 4 *If $t^C(p, w) \geq \underline{t}$, $(t^C(p, w), t^C(p, w))$ is a unique effort equilibrium.*

Table 1: The Agents' Best Response Functions Assuming Constant Rework Time

	Excess Capacity	Capacity Constrained
Self	$f^{-1}(\frac{1}{pr})$	$f^{-1}(\frac{1}{pr})$
Dedicated	$f^{-1}(\frac{\alpha}{pw})$	$f^{-1}(\frac{1-p\bar{F}(t^D)}{pt^D})$
Cross	$f^{-1}(\frac{\alpha}{pw})$	$f^{-1}(\frac{1-\rho(t^C)^2}{pt^C(1+\rho(t^C)(1-\frac{r}{t^C}))} - \frac{\bar{F}(t^C)}{t^C})$

Surprisingly, the agents' optimal effort in equilibrium is independent of each other's effort and is solely determined by the principal's quality inspection and incentive decisions. Because the agents have idle time in steady state, performing rework simply reduces idle time, but does not impact their workload of new jobs. Therefore, cross routing imposes no additional effect on the agents' incentives other than taking away the second opportunity to work on a job, the effect also present in dedicated routing. As a result, the agents have the same best response function as agent 1 in dedicated routing and thus have no strategic interactions. The principal's problem becomes

$$V^C = \max_{0 \leq p \leq 1, w \geq 0} 2\lambda[v - w - \bar{F}(t)(pc_I + (1-p)c_E)] - C_A(p), \quad (17)$$

$$\text{subject to} \quad \lambda[(1-p\bar{F}(t))w - \alpha t] \geq 0 \quad (\text{IR}), \quad (18)$$

$$t = t^C(p, w), \quad t \geq \underline{t} \quad (\text{IC}). \quad (19)$$

In the next subsection, we compare V^S , V^D , and V^C and identify which rework routing scheme achieves the highest profit rate.

4.5 Performance Comparison of Three Networks: Implicit Punishment

Comparing equation (3) with (8) allows us to illustrate the importance of the assumption (A1). Notice that when r is large, the difference between $f(t^{FB})$ and $f(t^S)$ becomes small and thus even the self routing scheme performs close to first best. This supports the intuition that agents have incentives to get it right in the first time when rework is costly. Therefore, assumption (A1) allows us to restrict our attention to the range of parameter values where agents' opportunistic behavior is prominent. For a summary of the best response functions of the agents, please refer to Table 1. To facilitate our comparison, we introduce a notation $t^{FB}(p)$ to represent the first-best solution at any fixed p . Further, let $w^* = \frac{\alpha}{p^{FB}f(t^{FB})}$, which we will show attains first best under certain conditions.

Proposition 1 *Self routing can never implement first best. Furthermore, $t^S(p) < t^{FB}(p)$ for all $p \in (0, 1)$. In contrast, dedicated routing implements first best with a unique contract $\{p^{FB}, w^*\}$ if and only if $w^* \geq \frac{\alpha t^{FB}}{1 - p^{FB} \bar{F}(t^{FB})}$, while cross routing implements first best with a unique contract $\{p^{FB}, w^*\}$ if and only if $w^* \geq \alpha[t^{FB} + p^{FB} \bar{F}(t^{FB})r]$.*

Proposition 1 reflects the weakness of the conventional self routing scheme: because the agent has a second chance to work on a job and earn the piece rate, he has a disincentive to exert effort in first pass and takes his chance at quality inspection. This gaming behavior of the agent leads to a lower first-pass quality level, incurring higher internal and external failure costs to the principal. In contrast, both dedicated and cross routing can attain first best under a mild condition on w^* .

From a central planner's point of view, dedicated and cross routing of rework are superior because the agent effort and quality inspection are set at the system optimal level. However, implementing $\{t^{FB}, p^{FB}\}$ in the dedicated and cross routing schemes does not guarantee that the IR constraints of the agents bind, i.e. the principal may need to pay a piece rate that leaves the agents with non-zero utility³. This implies that it may not be optimal for the principal to implement the effort-inducing contracts specified in Proposition 1. To address this issue, next we compare the three routing schemes based on the principal's profit rate. To this end, we first introduce three new notations. Let $w(p)$ be the optimal piece rate for the principal at any fixed p . This allows us to further define $t(p) := t(p, w(p))$ and quality output $Q(p) := Q(t(p))$.

Lemma 5 *For all $p \in (0, 1)$, $\min\{t^C(p), t^D(p)\} > t^S(p)$. Therefore, $\min\{Q^C(p), Q^D(p)\} > Q^S(p)$.*

We have shown previously that dedicated and cross routing share the same best response functions, which implies that the two schemes have the same incentive effects on effort. Lemma 5 further compares these two schemes with self routing in their ability to induce effort: dedicated and cross routing induce higher first-pass effort than self routing at any inspection precision, leading to a higher quality output. Dedicated and cross routing provide stronger incentives for quality because assigning rework to a different agent imposes an implicit punishment on the agents for their quality failure. This punishment is derived from the fact that in these two schemes the agents lose the effort spent on each job that fails quality inspection.

Lemma 6 *For all $p \in (0, 1)$, $\min\{w^C(p), w^D(p)\} > w^S(p)$.*

³If a lump sum transfer is allowed, the principal can extract all the surplus and let the agents earn zero utility in equilibrium.

Interestingly, we find that the piece rate paid in cross and dedicated routing are higher than the one paid in self routing because in the former two schemes the agents exert higher effort in equilibrium and cannot recoup the cost of effort spent on the jobs that have failed inspection. Lemmas 5 and 6 together highlight the principal's trade-off between inducing effort and bearing high compensation cost. The next proposition prescribes the routing scheme with the highest profit rate under different conditions⁴.

Proposition 2 *The rank order of the principal's profit rate depends on the quality costs and the disutility of effort:*

(a) $V^C \geq V^D$;

(b) *when the IR1 and the IR constraint bind under dedicated and cross routing, respectively,*

(i) *if c_I or c_E are sufficiently large, then $V^D > V^S$;*

(ii) *if α is sufficiently large, then $V^S > V^C$;*

(c) *when the IR1 and the IR constraint do not bind under dedicated and cross routing, respectively,*

(i) *if α is sufficiently large, then $V^S > V^C$.*

Proposition 2 states that the performance of the three networks crucially depends on the quality costs and the disutility of effort. When the internal and external failure costs are high, it is beneficial to induce effort to achieve higher first-pass quality. On the other hand, when the disutility of effort is high, the principal has to trade-off inducing higher first-pass effort with paying higher piece rates. Notice that the IR constraints of the dedicated and cross routing schemes may not bind because the principal's objective functions are not monotonically decreasing in w . When the IR constraints do not bind, we have to solve the principal's problem to derive the optimal w , which makes the solution analytically intractable without further assumption on the functional form of F . Only when α is sufficiently large can we determine a rank order of the three networks in terms of the principal's profit rate.

We now illustrate the comparison with an example. Since Proposition 2 shows that cross routing always weakly dominates dedicated routing, we will focus on the comparison between self

⁴The principal's problem maximizes a continuous function over a compact set, implying the existence of an optimum. Our comparison results do not require concavity of the objective function nor uniqueness of the optimum.

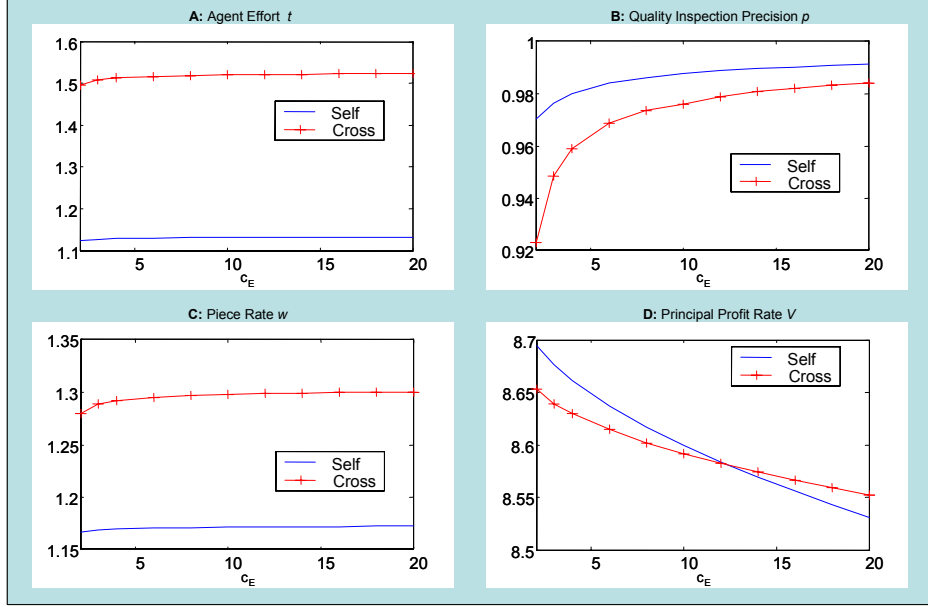


Figure 2: Equilibrium Performance Depends on the External Failure Cost: High Disutility of Effort ($\alpha = 0.8$). $F(t) = 1 - e^{-3(t-1)}$, $C_A(p) = \frac{0.001p^2}{1-p}$, $\lambda = 0.5$, $r = 0.5$, $c_I = 0.1$.

and cross routing. When the disutility of effort is high, Figure 2 shows that there exists a threshold of c_E , above which cross routing generates higher profit rate than self routing. However, when the disutility of effort is low, Figure 3 shows that cross routing always dominates self routing. In summary, when it is inexpensive for the principal to induce quality-improving effort, cross routing is better. Otherwise, it is only better for a range of c_E values up to a certain threshold.

5 Case II: Capacity Constrained

In contrast to the excess capacity case, the throughput, i.e., productivity, for a capacitated system is endogenous and depends on the agents' effort level. The higher the effort an agent exerts, i.e., the longer time he spends on each job, the lower his throughput becomes. Therefore, both the agents and the principal face a trade-off between throughput and quality. Since each agent is continuously busy⁵, the disutility of effort does not affect their best responses. Optimizing their utility rate, the agents balance the time allocated to new jobs vs. rework to trade-off earning the piece rate from first-pass success with that from rework. The principal balances inducing quality-improving effort with increasing productivity. Capacity constraint also impacts the stability condition of the

⁵The rework agent in dedicated routing has idle time in steady state.

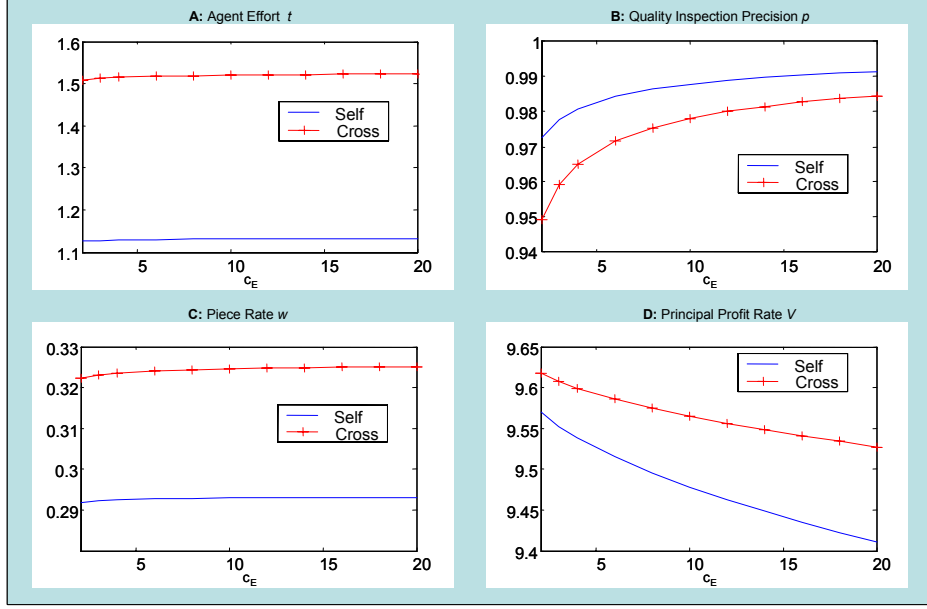


Figure 3: Equilibrium Performance Depends on the External Failure Cost: Low Disutility of Effort ($\alpha = 0.2$). $F(t) = 1 - e^{-3(t-1)}$, $C_A(p) = \frac{0.001p^2}{1-p}$, $\lambda = 0.5$, $r = 0.5$, $c_I = 0.1$.

system. We will use ρ_i to denote agent i 's utilization on rework. Since we focus on the dynamics of the rework queues, the stability condition is $\max_{i \in \{1,2\}} \rho_i < 1$.

5.1 The First-Best Benchmark

With capacity constraint, the agents are continuously busy and spend $t + p\bar{F}(t)r$ time units per job on average. Renewal theory yields that the agents' long-run average utility rate is $w/(t + p\bar{F}(t)r) - \alpha$. Under full information, the principal maximizes her long-run average profit rate:

$$V^{FB} = \max_{t \geq t, 0 \leq p \leq 1, w \geq 0} 2 \left[\frac{v - w - \bar{F}(t)(pc_I + (1-p)c_E) - C_A(p)}{t + p\bar{F}(t)r} \right] \quad (20)$$

$$\text{subject to} \quad \frac{w}{t + p\bar{F}(t)r} - \alpha \geq 0 \quad (\text{IR}). \quad (21)$$

Since V is monotonically decreasing in w , the IR constraint must bind and the optimization problem reduces to a two-variable problem of t and p .

Lemma 7 *Assume an interior first-best solution $\{t^{FB}, p^{FB}\}$ exists. Then it must satisfy*

$$f(t^{FB}) = \frac{1}{p^{FB}r + \frac{1}{A(t^{FB}, p^{FB})}(p^{FB}c_I + (1-p^{FB})c_E)}, \quad (22)$$

$$C'_A(p^{FB}) = \bar{F}(t^{FB})(c_E - c_I - A(t^{FB}, p^{FB})r), \quad (23)$$

where $A(t^{FB}, p^{FB}) = \frac{v - \bar{F}(t^{FB})(p^{FB}c_I + (1-p^{FB})c_E) - C_A(p^{FB})}{t^{FB} + p^{FB}\bar{F}(t^{FB})r}$.

Notice that the first-order conditions resemble Equations 3 and 4. The only difference is that the disutility of effort α is replaced by $A(t^{FB}, p^{FB})$.

5.2 Self Routing

When effort is not observable, the principal's objective function becomes

$$V^S = \max_{0 \leq p \leq 1, w \geq 0} 2 \left[\frac{v - w - \bar{F}(t)(pc_I + (1-p)c_E) - C_A(p)}{t + p\bar{F}(t)r} \right], \quad (24)$$

$$\text{subject to} \quad \frac{w}{t + p\bar{F}(t)r} - \alpha \geq 0 \quad (\text{IR}), \quad (25)$$

$$t \in \arg \max_{t' \geq \underline{t}} \left\{ \frac{w}{t' + p\bar{F}(t')r} - \alpha \right\} \quad (\text{IC}). \quad (26)$$

The first-order condition of the agents' problem is equivalent to $f(t^S) = \frac{1}{pr}$.

Lemma 8 *If $t^S \geq \underline{t}$, then t^S is a unique global maximum and is the agents' best response to the incentive contract $\{p, w\}$.*

Notice that the agents have the same best response function as in the excess capacity case. In both cases, the agents maximize their average payoff by minimizing the total expected time spent on each job, i.e.,

$$t^S = \arg \min_{t' \geq \underline{t}} \{t' + p\bar{F}(t')r\} \quad (27)$$

Doing so is optimal for the agents because the piece rate is a guaranteed income for each agent with the opportunity of rework. Consequently, the agents' optimal effort only depends on the inspection precision p and the slope of F , thus independent of whether the agents are continuously busy or have idle time.

5.3 Dedicated Routing

Without loss of generality, we assign new jobs to agent 1 and rework to agent 2. The principal maximizes

$$V^D = \max_{0 \leq p \leq 1, w \geq 0} \frac{1}{t_1} [v - w - \bar{F}(t_1)(pc_I + (1-p)c_E) - C_A(p)], \quad (28)$$

$$\text{subject to} \quad \frac{(1 - p\bar{F}(t_1))w}{t_1} - \alpha \geq 0 \quad (\text{IR1}), \quad \frac{w}{r} - \alpha \geq 0 \quad (\text{IR2}), \quad (29)$$

$$t_1 \in \arg \max_{t' \geq \underline{t}} \left\{ \frac{(1 - p\bar{F}(t'))w}{t'} - \alpha \right\} \quad (\text{IC1}). \quad (30)$$

Let t^D denote agent 1's best response. The first-order condition $U_1'(t) = 0$ is equivalent to

$$f(t^D) = \frac{1 - p\bar{F}(t^D)}{pt^D}. \quad (31)$$

Given agent 1's best response, agent 2's utilization $\rho_2 = \frac{p\bar{F}(t^D)r}{t^D}$. The stability condition $\rho_2 < 1$ is automatically satisfied because $r \leq t^D$ and $p < 1$.

Lemma 9 *If $t^D \geq \underline{t}$, then t^D is a unique global maximum and is agent 1's best response to the incentive contract $\{p, w\}$.*

Different from the excess capacity case, agent 1's optimal effort does not depend on w . Since agent 1 is continuously busy in the capacitated system, he does not face the trade-off between earning piece rate and having idle time. He only cares about the expected time spent on each job, and thus his successful throughput, represented by $(1 - p\bar{F}(t))/t$.

5.4 Cross Routing

Unlike in the case of excess capacity, we only need to consider the queuing dynamics of the two rework queues (because the new job queues are never empty when the system is capacitated). The state of the queuing network is described by (Z_3, Z_4) , where Z_{i+2} is the rework queue size for agent i . Figure 4C illustrates the state transitions of this multi-class queuing network. We define agent i 's rework utilization by $\rho_i(t_j) = p\bar{F}(t_j)\frac{r}{t_j}$, $i, j \in \{1, 2\}$, $i \neq j$. In steady state, $Z_3 Z_4 = 0$ holds because the states where both rework queues are nonempty are transient. Though we could have solved the limiting distribution for each possible state of (Z_3, Z_4) using the detailed balance equation approach, we only need the aggregate probabilities of the agents working on new jobs π_i^n and on rework π_i^r . In steady state, the queuing network must satisfy

$$\begin{aligned} \pi_i^n + \pi_i^r &= 1 \quad (\text{Law of total probability}), \\ \lambda_j \pi_j^n &= \mu_i^r \pi_i^r \quad (\text{Balance of agent } i \text{'s rework queue}), \end{aligned} \quad (32)$$

for $i, j \in \{1, 2\}$ and $i \neq j$. Solving the equations yields

$$\pi_i^n = \frac{1 - \rho_i}{1 - \rho_i \rho_j}, \quad \pi_i^r = \frac{\rho_i(1 - \rho_j)}{1 - \rho_i \rho_j}. \quad (33)$$

The stability condition is $\max\{\rho_1, \rho_2\} < 1$, which is automatically satisfied because $r \leq \min\{t_1, t_2\}$ and $p < 1$. Agent i 's long-run average utility rate

$$\begin{aligned} U_i(t_i, t_j) &= \pi_i^n \times \frac{w(1 - p\bar{F}(t_i)) - \alpha t_i}{t_i} + \pi_i^r \times \frac{w - \alpha r}{r} \\ &= \frac{w}{1 - \rho_i \rho_j} \left[\frac{(1 - p\bar{F}(t_i))(1 - \rho_i)}{t_i} + \frac{\rho_i(1 - \rho_j)}{r} \right] - \alpha \end{aligned} \quad (34)$$

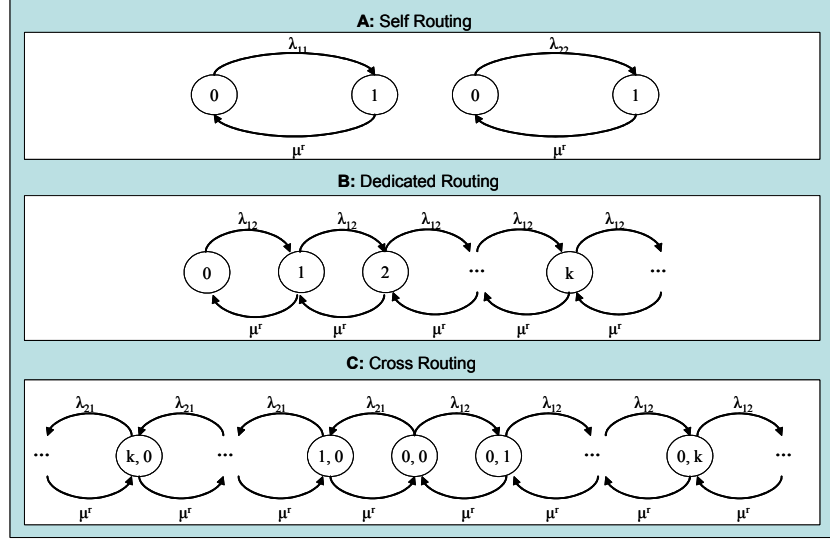


Figure 4: State Transition Diagrams of a Capacity Constrained Queuing System under the Three Rework Routing Schemes

Taking derivative of U_i w.r.t. t_i allows us to characterize the symmetric effort equilibrium given in Lemma 10.

Lemma 10 *There exists a symmetric effort equilibrium (t_i^C, t_j^C) with $t_i^C = t_j^C = t^C$, which must satisfy*

$$p[t^C f(t^C) + \bar{F}(t^C)][1 + \rho(t^C)(1 - \frac{r}{t^C})] + \rho(t^C)^2 - 1 = 0, \quad (35)$$

where $\rho(t^C) = p\bar{F}(t^C)(\frac{r}{t^C})$.

Similar to dedicated routing, the effort equilibrium only depends on p , i.e., $t^C = t^C(p)$. We now formulate the principal's objective function

$$\begin{aligned} V &= 2[\pi^n \times \frac{(v-w)(1-p\bar{F}(t)) + \bar{F}(t)(pc_I + (1-p)c_E) - C_A(p)}{t} + \pi^r \times \frac{v-w}{r}] \\ &= 2[\frac{(v-w)(1-p\bar{F}(t)) + \bar{F}(t)(pc_I + (1-p)c_E) - C_A(p)}{t + p\bar{F}(t)r}] \end{aligned} \quad (36)$$

Hence, the principal's problem becomes

$$V^C = \max_{w, 0 \leq p \leq 1} 2[\frac{(v-w)(1-p\bar{F}(t)) + \bar{F}(t)(pc_I + (1-p)c_E) - C_A(p)}{t + p\bar{F}(t)r}], \quad (37)$$

$$\text{subject to } \frac{w}{t + p\bar{F}(t)r} - \alpha \geq 0 \quad (\text{IR}), \quad (38)$$

$$t = t^C(p), \quad t \geq \underline{t} \quad (\text{IC}). \quad (39)$$

5.5 Comparing Three Networks: Externality and Prisoner's Dilemma

We compare the performance of the three networks. In Section 4, we have shown that dedicated and cross routing impose an implicit punishment for quality failure. Here, we will highlight two additional effects on the agents' effort in cross routing: externalities of workload allocation and a prisoner's dilemma.

Proposition 3 *The first-best solution $\{p^{FB}, t^{FB}\}$ can never be achieved by the three rework routing schemes. Furthermore, $t^S(p) < t^{FB}(p)$ for all $p \in (0, 1)$.*

The conventional self routing scheme induces lower effort than the first-best situation at any inspection precision p . As a result, self routing can never achieve first best. Dedicated and cross routing cannot attain first best either. Next we determine whether dedicated and cross routing can alleviate the moral hazard problem present in self routing.

Lemma 11 *For all $p \in (0, 1)$, $t^C(p) > t^D(p) > t^S(p)$ and therefore, $Q^C(p) > Q^D(p) > Q^S(p)$.*

Similar to the excess capacity case, self routing induces the least effort. However, cross routing induces even higher effort than dedicated routing. Under cross routing, the two parallel agents impact each other in two ways: they both generate and perform rework for each other. Since rework is favorable, each agent would like the other one to send him more rework. Because rework has priority, agent i has an incentive to pass less rework to agent j so that agent j has more time to work on new jobs and pass more rework back to agent i .

Externality. The strategic interaction in the effort game results in workload allocation externalities between the agents. Whenever agent i increases effort, he not only improves his first-pass success probability, but also forces agent j to spend more time on new jobs and thus generate more rework for agent i , keeping agent j 's effort unchanged. Analytically,

$$\frac{\partial \pi_j^n}{\partial t_i} = -\frac{1 - \rho_i}{(1 - \rho_i \rho_j)^2} \frac{\partial \rho_j}{\partial t_i} > 0, \quad \frac{\partial \pi_i^r}{\partial t_i} = -\frac{\rho_i(1 - \rho_i)}{(1 - \rho_i \rho_j)^2} \frac{\partial \rho_j}{\partial t_i} > 0. \quad (40)$$

$\partial \pi_j^n / \partial t_i > 0$ illustrates the workload externality imposed on agent j when agent i increases his first-pass effort. Since π_i^r is the fraction of time agent i spends on rework in steady state, $\partial \pi_i^r / \partial t_i > 0$ implies that agent i has more rework allocation when he increases his first-pass effort. For the same reason, agent j increases his first-pass effort to respond to agent i 's action. In the effort Nash equilibrium, both agents exert higher first-pass effort than in the dedicated routing scheme,

resulting in a lower defect rate. Therefore, the workload allocation externality in the effort game gives cross routing superiority in inducing quality-improving effort.

Lemma 12 For all $p \in (0, 1)$, $\min\{w^C(p), w^D(p)\} > w^S(p)$.

Lemma 12 states that the principal pays a higher piece rate to compensate for the higher effort that agents exert in cross and dedicated routing. More interestingly, using this piece rate ranking, we can show that the effort equilibrium arising in the cross routing scheme exhibits a prisoner's dilemma.

Prisoner's Dilemma. Notice that cooperative agents would exert t^S because it minimizes the total expected time spent on each job. This cooperative outcome gives agents strictly positive utility rate because $w^C(p) > w^S(p)$, thus a better outcome for both agents than the equilibrium outcome that renders zero utility rate for both agents. Since $f(t^S) = 1/pr$,

$$\begin{aligned} \frac{\partial U_i(t_i, t^S)}{\partial t_i} \Big|_{t_i=t^S} &= \frac{w(1 - \rho(t^S))}{[(1 - \rho(t^S)^2)t^S]^2} [p(t^S f(t^S) + \bar{F}(t^S))(1 + \rho(t^S)(1 - \frac{r}{t^S})) + \rho(t^S)^2 - 1] \\ &= \frac{w(1 - \rho(t^S))}{[(1 - \rho(t^S)^2)t^S]^2} [\frac{t^S}{r}(1 + \rho(t^S))(1 + \rho(t^S)(1 - \frac{r}{t^S})) + \rho(t^S)^2 - 1] \\ &> 0. \end{aligned} \tag{41}$$

The last inequality follows from the fact that $\rho(t^S) < 1$ and $r \leq t^S$. Therefore, agent i has an incentive to unilaterally deviate from the cooperative outcome. (Section 6.2 elaborates on this strategic behavior and discusses incentives for collusion.) This prisoner's dilemma works in favor of the principal because it induces higher first-pass effort and thus leads to higher quality output. We now compare the principal's profit rate in cross routing and self routing⁶.

Proposition 4 The rank order of the principal's profit rate depends on the quality costs and the gross margin:

(i) if c_I, c_E are sufficiently large or $C_A(\cdot)$ is sufficiently convex, $V^C > V^S$.

(ii) if v is sufficiently large, $V^S > V^C$.

⁶We shall not compare the profit rate of dedicated routing to the other two schemes because it has lower utilization of agent 2 by design. Notice that in dedicated routing, agent 2 is not continuously busy because the rework queue generated by agent 1 has positive probability of being empty under the stability condition of the system. In contrast, both agents are fully utilized in the other two schemes. This renders dedicated routing incomparable with the other two schemes.

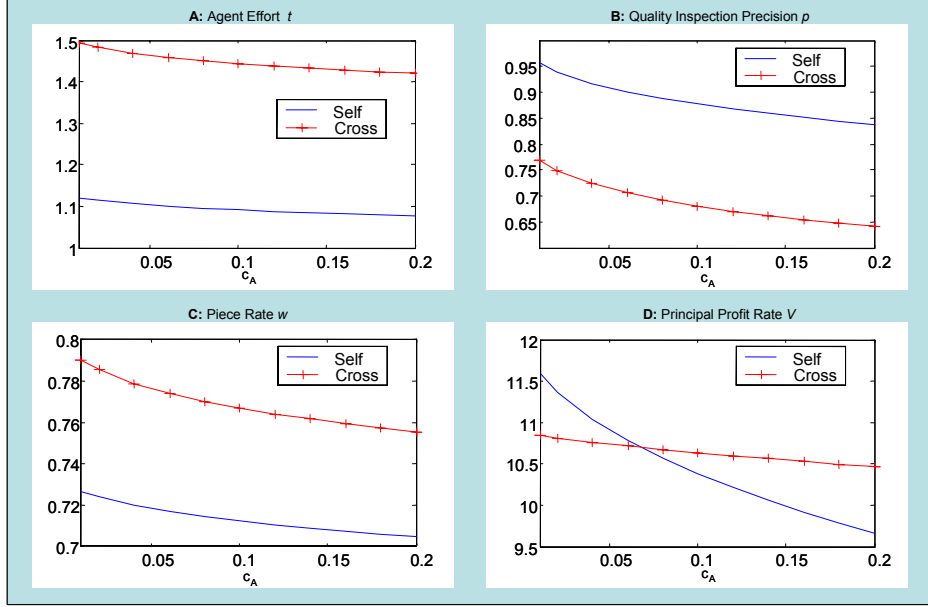


Figure 5: Equilibrium Performance Depends on the Appraisal Cost: High Gross Margin ($v = 10$). $F(t) = 1 - e^{-3(t-1)}$, $C_A(p) = \frac{c_A p^2}{1-p}$, $r = 0.5$, $\alpha = 0.5$, $c_I = 0.5$.

Being capacity constrained, the principal must take into account the effect of agents' effort on throughput. If she earns a high gross margin per job, the principal has less incentive to induce effort because high effort leads to low throughput and consequently lowers the revenue earned per unit time. Since cross routing induces high effort, the first-pass quality is improved at the expense of low throughput. Therefore, cross routing underperforms self routing when v is sufficiently large. However, when the costs of quality are high, it becomes critical for the principal to improve first-pass quality, making cross routing preferable to self routing. We illustrate these effects by numerical examples. When the gross margin is high (Figure 5), there exists a threshold of c_A (c_A indicates how convex $C_A(\cdot)$ is), above which cross routing generates higher profit rate than self routing. In contrast, when the gross margin is low (Figure 6), cross routing always dominates self routing.

6 Extensions

In this section, we discuss three extensions: the effect of non-constant rework time, prisoner's dilemma and incentives for collusion among agents, and monetary punishment as an alternative incentive contract.

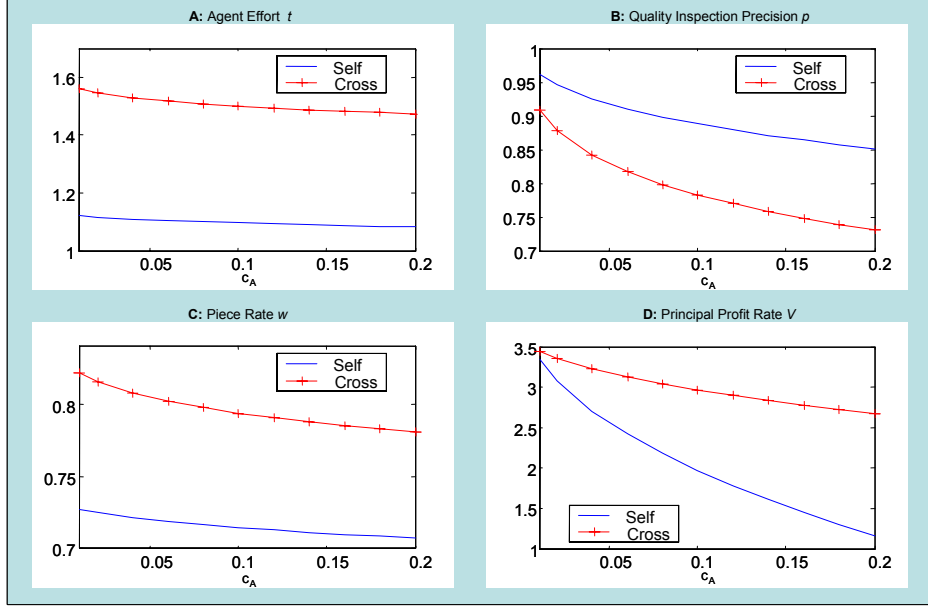


Figure 6: Equilibrium Performance Depends on the Appraisal Cost: Low Gross Margin ($v = 4$). $F(t) = 1 - e^{-3(t-1)}$, $C_A(p) = \frac{c_A p^2}{1-p}$, $r = 0.5$, $\alpha = 0.5$, $c_I = 0.5$.

6.1 Dependent Rework Time

We now relax the assumption that the rework time has a constant mean r . We let r depend on the first-pass effort t , i.e., $r = \tau - t$, where τ is a constant. While the agents' problems remain well behaved, the comparison of the three rework routing schemes becomes analytically less tractable. To verify that our main results presented earlier still hold with dependent rework time, we conducted an extensive numerical study. In this subsection, we will delve directly into the results without laying out the optimization problems of the principal and the agents. Note that these problems are identical to the ones presented earlier except that the rework time r is replaced with $\tau - t$ wherever appropriate. We summarize the agents' best response functions⁷ in Table 2.

Rearranging the best response function of the self routing scheme gives

$$\frac{\Pr(\text{First Pass})}{\Pr(\text{Failing Quality Inspection})} = \frac{1 - p\bar{F}(t^S)}{p\bar{F}(t^S)} = (\tau - t^S) \frac{f(t^S)}{\bar{F}(t^S)} \quad (42)$$

In words, the agents optimally set their effort level such that the ratio of passing vs. failing quality inspection is equal to the product of the average rework time and the hazard rate. Since in dedicated routing, rework is separated from new jobs, its optimal effort remains unchanged under the new

⁷For the capacity-constrained system under cross routing, the first-order condition is displayed in place of the best response function .

Table 2: The Agents' Best Response Functions Assuming Dependent Rework Time

	Excess Capacity	Capacity Constrained
Self	$f^{-1}\left(\frac{1-p\bar{F}(t^S)}{p(\tau-t^S)}\right)$	$f^{-1}\left(\frac{1-p\bar{F}(t^S)}{p(\tau-t^S)}\right)$
Dedicated	$f^{-1}\left(\frac{\alpha}{pw}\right)$	$f^{-1}\left(\frac{1-p\bar{F}(t^D)}{pt^D}\right)$
Cross	$f^{-1}\left(\frac{\alpha}{pw}\right)$	$\tau p \rho(t^C) [t^C f(t^C) + \bar{F}(t^C) - \frac{(t^C)^2}{\tau}] [\frac{1}{\tau-t^C} - \frac{1-p\bar{F}(t^C)}{t^C}] - (1 - \rho(t^C)^2) [1 - pt^C f(t^C) - p\bar{F}(t^C)] = 0$

assumption. The effort equilibrium induced in cross routing does not change in the case of excess capacity. In contrast, when the system is capacity constrained, the effort equilibrium in cross routing is determined by a complex equation. As a result, comparing performance analytically becomes very challenging. Nevertheless, our numerical results confirm our main results earlier. For example, Figure 7 shows the profit rate of the cross and self routing schemes under four different settings: (A) and (B) illustrate the cases of high and low disutility of effort, respectively, for a system with excess capacity; (C) and (D) illustrate the cases of high and low gross margin, respectively, for a system with capacity constraint.

6.2 Prisoner's Dilemma and Collusion

Until now, we have focused on a continuous work flow system where jobs are constantly assigned to agents. To highlight the prisoner's dilemma in cross routing, we now consider a project environment where a single job (i.e., project) is assigned to an agent who needs to complete it within certain period of time. It is more reasonable to assume that the agents maximize their utility per job in this context. For analytical purposes, we assume there are only two possible effort levels $\{t_H, t_L\}$. With effort levels t_H and t_L , good quality output is produced with probabilities π_H and π_L , respectively. We assume $0 < \pi_L < \pi_H < 1$. As in the main model, rework takes a constant effort r , which is assumed to be observable. Moreover, we assume rework is relatively less costly, specifically, $r < t_L$. Here we can allow a more general disutility⁸ of effort $g(t)$, with $g(0) = 0$, $g' > 0$ and $g'' > 0$. Further more, we assume that it is optimal for the principal to induce high effort. This is the more interesting case as quality is crucial to the principal.

Recall the intuition from Section 5 that cross routing has higher incentives for quality as a result of a prisoner's dilemma between the two parallel agents. In this subsection, we will be able

⁸In our main model, a linear disutility of effort is assumed because long-run analysis of the agents' utility rate requires additivity of disutility.

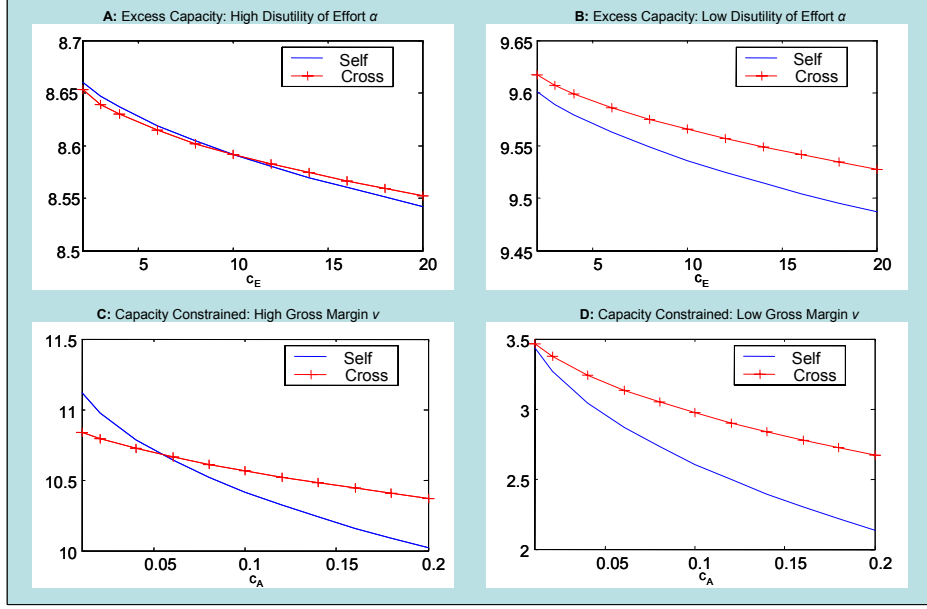


Figure 7: The Principal's Profit Rate in Systems with Excess Capacity (**A:** $\alpha = 0.8$. **B:** $\alpha = 0.2$. $\lambda = 0.5$, $c_A = 0.001$) and with Capacity Constraint (**C:** $v = 10$. **D:** $v = 4$. $c_E = 10$). $F(t) = 1 - e^{-3(t-1)}$, $C_A(p) = \frac{c_A p^2}{1-p}$, $\tau = 2$, $c_I = 0.5$.

to demonstrate this effect more directly.

6.2.1 Self Routing

In self routing, agent i 's utility depends on his effort:

$$U_H = w - g(t_H) - p(1 - \pi_H)g(r), \quad U_L = w - g(t_L) - p(1 - \pi_L)g(r). \quad (43)$$

The IC constraint for inducing high effort is $U_H \geq U_L$. Equivalently,

$$p \geq \bar{p}^S = \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)g(r)}. \quad (44)$$

To ensure high effort is implementable using the self routing scheme, we need $\bar{p}^S < 1$ and thus assume $g(r) > \frac{g(t_H) - g(t_L)}{\pi_H - \pi_L}$. This is the more interesting case because the self routing scheme would otherwise be immediately inferior. Since the throughput is limited to one unit, the principal's problem becomes a cost minimization:

$$C^S = \min_{0 \leq p \leq 1, w \geq 0} w + (1 - \pi_H)(pc_I + (1 - p)c_E) + C_A(p) \quad (45)$$

$$\text{subject to} \quad w - g(t_H) - p(1 - \pi_H)g(r) \geq 0 \quad (\text{IR}), \quad p \geq \bar{p}^S \quad (\text{IC}). \quad (46)$$

6.2.2 Dedicated Routing

In dedicated routing, agent 1's utility depends on his effort:

$$U_H = (1 - p(1 - \pi_H))w - g(t_H), \quad U_L = (1 - p(1 - \pi_L))w - g(t_L). \quad (47)$$

The IC constraint for inducing high effort is $U_H \geq U_L$. Equivalently,

$$p \geq \bar{p}^D = \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)w}. \quad (48)$$

The principal's problem becomes

$$C^D = \min_{0 \leq p \leq 1, w \geq 0} w + (1 - \pi_H)(pc_I + (1 - p)c_E) + C_A(p) \quad (49)$$

$$\text{subject to} \quad (1 - p(1 - \pi_H))w - g(t_H) \geq 0 \quad (\text{IR1}), \quad p \geq \bar{p}^D \quad (\text{IC1}), \quad (50)$$

$$p(1 - \pi_H)(w - g(r)) \geq 0 \quad (\text{IR2}). \quad (51)$$

Since \bar{p}^D is dependent on w , high effort is always implementable as long as the piece rate w is set high enough. Similar to the excess capacity case of the main model, dedicated routing gives the principal an additional lever, i.e., w , to induce quality-improving effort.

6.2.3 Cross Routing

In cross routing, agent i 's utility depends on agent j 's effort:

$$\begin{aligned} U_{HH} &= w - g(t_H) - p(1 - \pi_H)g(r), & U_{LH} &= (1 - p(\pi_H - \pi_L))w - g(t_L) - p(1 - \pi_H)g(r), \\ U_{LL} &= w - g(t_L) - p(1 - \pi_L)g(r), & U_{HL} &= (1 + p(\pi_H - \pi_L))w - g(t_H) - p(1 - \pi_L)g(r), \end{aligned} \quad (52)$$

where the first and second subscripts denote agent i 's and j 's effort level, respectively. The IC constraint for inducing $\{H, H\}$ equilibrium outcome is $U_{HH} \geq U_{LH}$. Equivalently,

$$p \geq \bar{p}^C = \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)w} \quad (53)$$

Similar to dedicated routing, high effort is always implementable as long as the piece rate w is set high enough. The principal's problem becomes

$$C^C = \min_{0 \leq p \leq 1, w \geq 0} w + (1 - \pi_H)(pc_I + (1 - p)c_E) + C_A(p) \quad (54)$$

$$\text{subject to} \quad w - g(t_H) - p(1 - \pi_H)g(r) \geq 0 \quad (\text{IR}), \quad p \geq \bar{p}^C \quad (\text{IC}). \quad (55)$$

6.2.4 Comparing Three Schemes

We first compare the lower bound on the quality inspection precision required to achieve high effort. It is simple to show that $\bar{p}^C = \bar{p}^D < \bar{p}^S$. The inequality follows from $w > g(r)$, which is true because of the IR constraint and the assumption that $r < \underline{t}$. Further notice that when $\bar{p}^C < p < \bar{p}^S$, the Nash equilibrium induced between the two agents in cross routing exhibits a prisoner's dilemma. This follows from

$$U_{HL} - U_{LL} = p(\pi_H - \pi_L)w - g(t_H) + g(t_L) > 0 \quad (56)$$

$$U_{HH} - U_{LL} = g(t_L) - g(t_H) + pg(r)(\pi_H - \pi_L) < 0 \quad (57)$$

In other words, even though strategy profile $\{L, L\}$ let both agents enjoy a higher payoff than the equilibrium payoff, agents will make unilateral deviation to high effort, resulting in a prisoner's dilemma. Interestingly, the existence of the prisoner's dilemma hinges on the condition that $p < \bar{p}^S$. If to the contrary $p \geq \bar{p}^S$, then $U_{HH} \geq U_{LL}$ and thus cross routing and self routing have the same incentive effects. Therefore, cross routing has stronger incentives for quality only when the optimal p under this scheme is strictly less than \bar{p}^S .

The equality between \bar{p}^C and \bar{p}^D implies that cross assignment of rework and assigning rework to a dedicated agent have equivalent effects on the agents' effort. In cross routing, though the agents could have exhibit more opportunistic behavior, it is curbed by the prisoner's dilemma. In dedicated routing, separating the rework task completely from the new job task alleviates the moral hazard problem. Consistent with the result of the main model, self routing is disadvantageous in inducing effort as it requires a higher inspection precision.

Proposition 5 *The rank order of the principal's cost depends on the quality costs:*

- (i) *if c_I is sufficiently large or $C_A(\cdot)$ is sufficiently convex, then $C^C = C^D < C^S$;*
- (ii) *if c_E is sufficiently large, then $C^C = C^D = C^S$.*

The above results differ from the main results (i.e., Propositions 2 and 4) in two important ways: (1) self routing is weakly dominated by the other two schemes; (2) c_I and c_E play different roles in determining the rank order. These differences result from the assumption that high effort is always desirable in the current model, while the principal is allowed to choose optimal effort in the main model.

Finally, we caution that the superior performance of cross routing relies on the restriction that the agents do not have future interactions. In a repeated game, a collusive outcome $\{L, L\}$ can be supported with Nash reversion strategies and a sufficiently large discount factor. This suggests that in practice, it may be beneficial for the principal to maintain a certain level of staff turnover to prevent collusion.

6.3 Punishment as an Alternative Incentive Scheme

Assigning rework to a different agent implicitly punishes the agent for shirking. In dedicated and cross routing, the agents are punished because they cannot recoup the cost of effort spent on a job that fails quality inspection. Such punishment could be replicated by a modified self routing scheme where the principal executes a monetary punishment whenever a defect is identified. Consider the case of excess capacity. Suppose the principal specifies a penalty x for each defect identified, the agents' problem becomes

$$\max_{t \geq \bar{t}} \lambda[w - \alpha(t + p\bar{F}(t)r) - p\bar{F}(t)x]. \quad (58)$$

The first-order condition is equivalent to $f(t) = \frac{1}{pr + \frac{1}{\alpha}px}$. Recalling Equation (3), we set $x = c_I + \frac{1-p^{FB}}{p^{FB}}c_E$ to allow the principal to achieve the first-best effort level. Different from the first-best outcome, the principal needs to compensate the agent with a higher piece rate: $w = w^{FB} + p^{FB}\bar{F}(t^{FB})x > w^{FB}$. Similarly, we could also design a piece rate plus a bonus that is paid whenever a job passes quality inspection in the first pass. This contract also enables the principal to achieve first best, but requires a higher piece rate than w^{FB} as well. Since positive rents need to be paid to the agents to attain t^{FB} and p^{FB} , it may not be profitable for the principal to implement such contracts.

While these incentive schemes are powerful, it is difficult in practice for a principal to “force” rework without or with negative pay. Moreover, the principal may face a limited liability constraint that withholds him from using the punishment compensation scheme⁹. In contrast, cross routing of rework is a more “fair” contract in the sense that the principal always pays the full piece rate but chooses to pay the parallel agent for his rework. Rather than designing a more complex contract, we take piece rate, a commonly used incentive scheme, as given, and try to design rework routing schemes that improve quality. We believe doing so has rather practical implications as the

⁹For a detailed discussion of limited liability constraint, see Sappington (1983) that illustrates limited liability in a risk-neutral agent setting.

cross routing scheme can be implemented without modifying the existing pay scheme, though the magnitude of piece rate may be changed.

7 Conclusions

This paper investigates how incentives and judicious rework routing can improve quality and profitability of a firm using a principal-agent model integrated into a multi-class queuing network. We demonstrate that conventional self routing of rework is always suboptimal in terms of inducing quality-improving effort. In contrast, dedicated and cross routing perform better in inducing effort. However, the principal’s financial performance depends not only on the first-pass effort induced, but also on quality-related costs, revenues, and agent disutilities of effort. The more novel cross routing scheme is applicable in both manufacturing and service operations environment. The merit of this routing scheme lies in the fact that the two parallel agents influence each other’s workload allocation over new jobs and rework in a way that leads to higher equilibrium first-pass effort as a result of a prisoner’s dilemma. This works in favor of the principal when quality is important, i.e., when quality costs are high.

We have made two methodological contributions to the agency and operations management literature. First, we study a multi-agent principal-agent model in a multi-class queuing network with endogenous queues (recall the job arrival rate of the rework queues is endogenously determined by the agents’ first-pass effort). To the best of our knowledge, this is the first attempt at modeling endogenous queuing dynamics in a principal-agent framework¹⁰. Second, we embed the quantity-quality trade-off in one decision variable, i.e., the processing time per job (or effort level). We think this is a more realistic way to model the trade-off because quantity and quality output are often inseparable tasks in a worker’s job.

There are three main limitations of the model. First, due to the inherent variability in queuing networks, risk aversion of agents cannot be easily incorporated given that we conduct long-run analysis. Second, we assume that agents commit to a single first-pass effort level even though in reality they can adjust effort from time to time and thus play a repeated game. Third, our model does not capture customer waiting costs and inventory holding costs, though they can be incorporated. When customer waiting costs are considered, pricing of the goods or services sold by

¹⁰Gilbert & Weng (1998) model a two-server queuing network in a principal-agent framework, but only consider queues with exogenous arrival rate.

the principal will depend on the agents' effort level. Customer waiting also affects the principal's decision on capacity, i.e., whether to acquire adequate staffing to provide good service or maintain high utilization of resources to minimize cost. Inventory holding costs can be incorporated straightforwardly. We believe this will change our result in one direction: the principal will have less incentives to induce effort because higher first-pass effort leads to longer flow time, and thus higher inventory holding costs.

Acknowledgements

We thank Kevin Wilson (Memphis Auto Auction) for bringing cross routing to our attention. We also greatly benefited from the encouraging suggestions from James Dana and Martin Lariviere.

References

- Afeche, P. & Mendelson, H. (2004), 'Pricing and priority auctions in queueing systems with a generalized delay cost structure', *Management Science* **50**(7), 869–882.
- Avriel, M., Diewert, W. E., Schaible, S. & Zang, I. (1988), *Generalized Concavity*, Plenum Press, New York.
- Baiman, S., Fischer, P. E. & Rajan, M. V. (2000), 'Information, contracting, and quality costs', *Management Science* **46**(6), 776–789.
- Cachon, G. & Harker, P. T. (2002), 'Competition and outsourcing with scale economies', *Management Science* **48**(10), 1314–1333.
- Gilbert, S. M. & Weng, Z. K. (1998), 'Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective', *Management Science* **44**(12), 1662–1669.
- Gunes, E. D. & Aksin, O. Z. (2004), 'Value creation in service delivery: Relating market segmentation, incentives, and operational performance', *Manufacturing and Service Operations Management* **6**(4), 338–357.
- Ha, A. Y. (2001), 'Optimal pricing that coordinates queues with customer-chosen requirements', *Management Science* **47**(7), 915–930.
- Hamilton, B. H., Nickerson, J. A. & Owan, H. (2003), 'Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation', *Journal of Political Economy* **111**(3), 465–497.
- Holmstrom, B. & Milgrom, P. (1991), 'Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design', *Journal of Law, Economics, and Organization* **7**, 24–52.

- Juran, J. M. & Gryna, F. M. (1993), *Quality Planning and Analysis: from product development through use*, McGraw-Hill, New York.
- Lazear, E. P. (2000), ‘Performance and productivity’, *American Economic Review* **90**(5), 1346–1361.
- Mendelson, H. & Whang, S. (1990), ‘Optimal incentive-compatible priority pricing for the m/m/1 queue’, *Operations Research* **38**(5), 870–883.
- Naor, P. (1969), ‘The regulation of queue size by levying tolls’, **37**, 15–24.
- Parlakturk, A. K. & Kumar, S. (2004), ‘Self-intereted routing in queueing networks’, *Management Science* **50**(7), 949–966.
- Plambeck, E. L. & Zenios, S. A. (2000), ‘Performance-based incentives in a dynamic principal-agent model’, *Manufacturing and Service Operations Management* **2**(3), 240–263.
- Reynier, D. J. & Tapiero, C. S. (1995), ‘The delivery and control of quality in supplier-producer contracts’, *Management Science* **41**(10), 1581–1589.
- Sappinton, D. (1983), ‘Limited liability contracts between principal and agent’, *Journal of Economic Theory* **29**, 1–21.
- Shumsky, R. A. & Pinker, E. J. (2003), ‘Gatekeepers and referrals in services’, *Management Science* **49**(7), 839–856.
- Van Mieghem, J. A. (2000), ‘Price and service discrimination in queuing systems: Incentive compatibility of gcu scheduling’, *Management Science* **46**(9), 1249–1267.

Appendix

PROOF OF LEMMA 1. We only need to show that the optimum exists and is an interior solution. Notice that when $t \geq \frac{v}{\alpha}$, $V < 0$. Therefore, we are maximizing a continuous function over a compact set: $t^{FB} \in [\underline{t}, \frac{v}{\alpha}]$ and $p^{FB} \in [0, 1]$, which implies the existence of an optimum. Moreover, $C'_A(1) = \infty$ implies $p^{FB} < 1$. $V(\frac{v}{\alpha}, p) < 0$ implies that $t^{FB} < \frac{v}{\alpha}$. Since the IR constraint must bind, substitute $w = \alpha(t + p\bar{F}(t)r)$ into the objective function and derive the Kuhn-Tucker conditions:

$$\begin{aligned} \frac{\partial V(t^{FB}, p^{FB})}{\partial t} &= 2\lambda[-\alpha(1 - p^{FB}f(t^{FB})r) + f(t^{FB})(p^{FB}c_I + (1 - p^{FB})c_E)] \\ &= 2\lambda[-\alpha + f(t^{FB})(p^{FB}(\alpha r + c_I) + (1 - p^{FB})c_E)] \leq 0, \end{aligned} \quad (59)$$

$$\frac{\partial V(t^{FB}, p^{FB})}{\partial p} = 2\lambda[\bar{F}(t^{FB})(c_E - c_I - \alpha r) - C'_A(p^{FB})] \leq 0. \quad (60)$$

Since $C'_A(0) = 0$, we must have $p^{FB} > 0$ because the assumption that $c_E - c_I > \alpha r$ leads to $\frac{\partial V(t^{FB}, 0)}{\partial p} > 0$. $p^{FB} \in (0, 1)$ implies that $(p^{FB}(\alpha r + c_I) + (1 - p^{FB})c_E) \in (\alpha r + c_I, c_E)$. If $\frac{\alpha}{f(\underline{t})} < \alpha r + c_I$, then $\frac{\partial V(\underline{t}, p)}{\partial t} > 0$ for any $p \in (0, 1)$. Therefore, $t^{FB} > \underline{t}$. ■

PROOF OF LEMMA 2. The second-order condition (SOC) is $\lambda \alpha p r F''(t) < 0$. ■

PROOF OF LEMMA 3. The SOC is $2\lambda w p F''(t) < 0$. ■

PROOF OF LEMMA 4. The SOC is $\lambda w p F''(t) < 0$. Therefore, t^C uniquely maximizes both agents' utility rate and thus (t^C, t^C) is a unique effort equilibrium. ■

PROOF OF PROPOSITION 1. (Self Routing) Substituting $t^S(p)$ into the FOC of t^{FB} gives $2\lambda[-\alpha(1 - p)f(t^S(p))r + f(t^S(p))(pc_I + (1 - p)c_E)] = 2\lambda[f(t^S(p))(pc_I + (1 - p)c_E)] > 0$. Since $\frac{\partial^2 V(t, p)}{\partial t^2} = 2\lambda F''(t)(p(\alpha r + c_I) + (1 - p)c_E) < 0$ for the first-best problem, it follows that $t^S(p) < t^{FB}(p)$ for all $p \in (0, 1)$. Therefore, first best can never be achieved. (Dedicated Routing) Since $t^D = f^{-1}(\frac{\alpha}{wp})$, the principal can set $p = p^{FB}$ and $w^* = \frac{\alpha}{p^{FB} f(t^{FB})}$ to induce t^{FB} . Further, if $w^* \geq \frac{\alpha t^{FB}}{1 - p^{FB} F(t^{FB})}$, then the IR1 and IR2 constraints are satisfied. The “only if” part is immediate because the IR1 constraint must be satisfied to implement the first-best solution. Similarly we prove the results for cross routing. ■

PROOF OF LEMMA 5. The IR1 constraint requires that $w(p) > \alpha \underline{t} \geq \alpha r$, which implies $\frac{\alpha}{pw(p)} < \frac{1}{pr}$. Since $F''(\cdot) < 0$, it follows that $t^D(p) = f^{-1}(\frac{\alpha}{pw(p)}) > f^{-1}(\frac{1}{pr}) = t^S(p)$. Similarly, $t^C(p) > t^S(p)$. ■

PROOF OF LEMMA 6. Since the principal's objective function under self routing is monotonically decreasing in w , the IR constraint must bind, i.e., $w^S(p) = \alpha[t^S(p) + pr\bar{F}(t^S(p))]$. However, $w^C(p) \geq \alpha[t^C(p) + pr\bar{F}(t^C(p))]$. It follows that

$$\begin{aligned} w^C(p) - w^S(p) &\geq \alpha[t^C(p) - t^S(p) + pr(\bar{F}(t^C(p)) - \bar{F}(t^S(p)))] \\ &= \alpha(t^C(p) - t^S(p))(1 - pr \frac{F(t^C(p)) - F(t^S(p))}{t^C(p) - t^S(p)}) \\ &> \alpha(t^C(p) - t^S(p))(1 - pr f(t^S(p))) = 0. \end{aligned} \quad (61)$$

The last inequality follows from the fact that $F(\cdot)$ is strictly concave and $t^C(p) > t^S(p)$. Since $w^D(p) \geq \frac{\alpha t^D(p)}{1 - p\bar{F}(t^D(p))}$,

$$\begin{aligned} w^D(p) - w^S(p) &\geq \frac{\alpha t^D(p)}{1 - p\bar{F}(t^D(p))} - \alpha[t^S(p) + pr\bar{F}(t^S(p))] \\ &> \frac{\alpha t^D(p)}{1 - p\bar{F}(t^D(p))} - \alpha[t^D(p) + pr\bar{F}(t^D(p))] \\ &= \frac{\alpha}{1 - p\bar{F}(t^D(p))} p\bar{F}(t^D(p))[t^D(p) - (1 - p\bar{F}(t^D(p)))r] > 0. \end{aligned} \quad (62)$$

The second inequality follows from the fact that $t^S = \arg \min_t \{t + pr\bar{F}(t)\}$. ■

PROOF OF PROPOSITION 2. Let p^S denote the optimal p under self routing. Similarly, we define p^D and p^C . Since $p \in [0, 1]$ and $w \in [0, v]$, we are maximizing a continuous function over a compact set, implying existence of an optimum. (Part a) Because the IR2 constraint of dedicated routing is always nonbinding and $f^D(t) = f^S(t) = \frac{\alpha}{pw}$, the principal's problems of dedicated and cross routing are identical except that the IR constraints are different. Under dedicated routing $w \geq \frac{\alpha t}{1 - p\bar{F}(t)}$, while under cross routing $w \geq \alpha(t + p\bar{F}(t)r)$. For any $t \geq \underline{t}$,

$$\frac{\alpha t}{1 - p\bar{F}(t)} - \alpha(t + p\bar{F}(t)r) = \alpha \frac{p\bar{F}(t)[t - (1 - p\bar{F}(t))r]}{1 - p\bar{F}(t)} > 0. \quad (63)$$

Therefore, the optimization problem in the dedicated routing has a more stringent IR constraint. Hence, $V^C \geq V^D$.

(Part b) We start by comparing the principal's profit rate at any p ,

$$\begin{aligned} V^C(p) - V^S(p) &= \lambda[-\alpha((t^C(p) + p\bar{F}(t^S(p)))r) + \alpha(t^S(p) + p\bar{F}(t^S(p)))r] \\ &\quad + (pc_I + (1 - p)c_E)(\bar{F}(t^C(p)) - \bar{F}(t^S(p))) \\ &= \lambda[-\alpha(t^C(p) - t^S(p))(1 - pr \frac{F(t^C(p)) - F(t^S(p))}{t^C(p) - t^S(p)}) \\ &\quad + (pc_I + (1 - p)c_E)(F(t^C(p)) - F(t^S(p)))] \end{aligned} \quad (64)$$

Since $\frac{F(t^C(p)) - F(t^S(p))}{t^C(p) - t^S(p)} < f(t^S(p)) = \frac{1}{pr}$, $1 - pr \frac{F(t^C(p)) - F(t^S(p))}{t^C(p) - t^S(p)} > 0$. Therefore $V^C(p) < V^S(p)$ if α is sufficiently large. Hence, $V^C = V^C(p^C) < V^S(p^C) < V^S(p^S) = V^S$ if α is sufficiently large. The last inequality follows from the optimality of p^S .

$$\begin{aligned} V^D(p) - V^S(p) &= \lambda[-\alpha(\frac{t^D(p)}{1 - p\bar{F}(t^D)} - (t^S(p) + p\bar{F}(t^S(p)))r) \\ &\quad + (pc_I + (1 - p)c_E)(\bar{F}(t^D(p)) - \bar{F}(t^S(p)))] \end{aligned} \quad (65)$$

Since $\frac{t^D(p)}{1 - p\bar{F}(t^D)} > t^D(p) + p\bar{F}(t^D)r > t^S(p) + p\bar{F}(t^S(p))r$, $V^D(p) > V^S(p)$ if c_I and c_E are sufficiently large. Using a similar argument as before, we prove that $V^D > V^S$ if c_I and c_E are sufficiently large.

(Part c) Since $w^C(p) > \alpha[t^C(p) + p\bar{F}(t^C(p))r]$

$$\begin{aligned} V^C(p) - V^S(p) &< \lambda[-\alpha(t^C(p) - t^S(p))(1 - pr \frac{F(t^C(p)) - F(t^S(p))}{t^C(p) - t^S(p)}) \\ &\quad + (pc_I + (1 - p)c_E)(F(t^C(p)) - F(t^S(p)))] \\ &< 0 \text{ if } \alpha \text{ is sufficiently large.} \end{aligned} \quad (66)$$

Using a similar argument as before, we prove that $V^C < V^S$ if α is sufficiently large. ■

PROOF OF LEMMA 7. Since $t^{FB} \in [\underline{t}, \frac{v}{\alpha}]$ and $p^{FB} \in [0, 1]$, we are maximizing a continuous function over a compact set, implying the existence of an optimum. Assuming an interior optimum exists, optimum $\{t^{FB}, p^{FB}\}$ is then given by the FOCs. ■

PROOF OF LEMMA 8. Evaluating the second derivative of $U(t)$ at t^S using $f(t^S) = 1/pr$ yields that

$$U''(t^S) = \frac{wprF''(t^S)}{(t^S + p\bar{F}(t^S)r)^2} + \frac{2w(1 - prf(t^S))}{(t^S + p\bar{F}(t^S)r)^3} = \frac{wprF''(t^S)}{(t^S + p\bar{F}(t^S)r)^2} < 0. \quad (67)$$

Because $U(t)$ is strictly concave at any interior critical point, $U(t)$ is strictly pseudoconcave (Avriel, Diewert, Schaible & Zang (1988)) and thus t^S is a unique global maximum if $t^S \geq \underline{t}$. ■

PROOF OF LEMMA 9. Evaluating the second derivative of $U_1(t_1)$ at t^D using $f(t^D) = \frac{1 - p\bar{F}(t^D)}{pt^D}$ yields that

$$U_1''(t^D) = w\left[\frac{pF''(t^D)}{t^D} - \frac{2}{(t^D)^3}(pt^D f(t^D) + p\bar{F}(t^D) - 1)\right] = \frac{wpF''(t^D)}{t^D} < 0. \quad (68)$$

Because $U_1(t_1)$ is strictly concave at any interior critical point, $U_1(t_1)$ is strictly pseudoconcave (Avriel et al. (1988)) and thus t^D is a unique global maximum if $t^D \geq \underline{t}$.

PROOF OF LEMMA 10.

$$\frac{\partial U_i(t_i, t_j)}{\partial t_i} = \frac{w(1 - \rho_i)}{(1 - \rho_i\rho_j)^2} \frac{1}{t_i^2} [p(t_i f(t_i) + \bar{F}(t_i))(1 + \rho_i(1 - \frac{r}{t_i})) + \rho_i\rho_j - 1]. \quad (69)$$

Equivalently,

$$p[t_i f(t_i) + \bar{F}(t_i)][1 + \rho_i(1 - \frac{r}{t_i})] + \rho_i\rho_j(t_i) - 1 = 0. \quad (70)$$

Let \hat{t}_i be the critical point satisfying the above FOC. The SOC evaluated at \hat{t}_i is

$$\begin{aligned} \frac{\partial^2 U_i(t_i, t_j)}{\partial t_i^2} \Big|_{t_i=\hat{t}_i} &= \frac{w(1 - \rho_i)}{(1 - \rho_i\rho_j(\hat{t}_i))^2} \frac{1}{\hat{t}_i^2} [p\hat{t}_i F''(\hat{t}_i)(1 + \rho_i(1 - \frac{r}{\hat{t}_i})) \\ &+ p(\hat{t}_i f(\hat{t}_i) + \bar{F}(\hat{t}_i)) \frac{\rho_i r}{\hat{t}_i^2} + \rho_i \frac{\partial \rho_j(\hat{t}_i)}{\partial t_i}] \\ &+ w(1 - \rho_i) \frac{\partial [\frac{1}{(1 - \rho_i\rho_j)^2} \frac{1}{t_i^2}]}{\partial t_i} [p(\hat{t}_i f(\hat{t}_i) + \bar{F}(\hat{t}_i))(1 + \rho_i(1 - \frac{r}{\hat{t}_i})) + \rho_i\rho_j(\hat{t}_i) - 1] \end{aligned} \quad (71)$$

Substituting $\frac{\partial \rho_j(\hat{t}_i)}{\partial t_i} = -\frac{pr}{\hat{t}_i^2} [t_i f(\hat{t}_i) + \bar{F}(\hat{t}_i)]$ and the FOC into the SOC gives

$$\frac{\partial^2 U_i(\hat{t}_i, t_j)}{\partial t_i^2} = \frac{w(1 - \rho_i)}{(1 - \rho_i\rho_j(\hat{t}_i))^2} \frac{1}{\hat{t}_i^2} [p\hat{t}_i F''(\hat{t}_i)(1 + \rho_i(1 - \frac{r}{\hat{t}_i}))] < 0 \quad (72)$$

The inequality follows from the fact that $F''(\cdot) < 0$ and that $(1 + \rho_i(1 - \frac{r}{\hat{t}_i})) > 0$. Because it is strictly concave at any interior critical point, $U_i(t_i, t_j)$ is strictly pseudoconcave in t_i (Avriel et al. (1988)), implying

\hat{t}_i is a unique global maximum. If we further assume a symmetric equilibrium, i.e., $\hat{t}_i = \hat{t}_j = t^C$, then we get the equilibrium equation. ■

PROOF OF PROPOSITION 3. Let $A(t, p) = \frac{v-w-\bar{F}(t)(pc_I+(1-p)c_E)-C_A(p)}{t+p\bar{F}(t)r}$ and $B(t, p) = t+p\bar{F}(t)r$.

Substituting $t^S(p)$ into the first-best FOC of $t^{FB}(p)$ gives

$$\begin{aligned} & \frac{f(t^S(p))(pc_I + (1-p)c_E)}{B(t^S(p), p)} - (1 - prf(t^S(p))) \frac{A(p, t^S(p))}{B(p, t^S(p))} \\ &= \frac{f(t^S(p))(pc_I + (1-p)c_E)}{B(t^S(p), p)} > 0 \end{aligned} \quad (73)$$

Claim: At any fixed p , V^{FB} is uniquely maximized at t^{FB} .

$$\begin{aligned} \frac{\partial V^2(t^{FB})}{\partial t^2} &= \frac{F''(t^{FB})(pc_I + (1-p)c_E)}{B(t^{FB}, p)} - (1 - prf(t^{FB})) \frac{f(t^{FB})(pc_I + (1-p)c_E)}{B(t^{FB}, p)^2} \\ &+ prF''(t^{FB}) \frac{A(t^{FB}, p)}{B(t^{FB}, p)} - (1 - prf(t^{FB})) \frac{f(t)(pc_I + (1-p)c_E)}{B(p, t^{FB})^2} \\ &+ 2[1 - prf(t^{FB})]^2 \frac{A(t^{FB}, p)}{B(t^{FB}, p)^2} \\ &= \frac{F''(t^{FB})(pc_I + (1-p)c_E)}{B(t^{FB}, p)} + prF''(t^{FB}) \frac{A(t^{FB}, p)}{B(t^{FB}, p)^2} \\ &- 2 \frac{(1 - prf(t^{FB}))}{B(t^{FB}, p)^2} [f(t)(pc_I + (1-p)c_E) - (1 - prf(t^{FB}))A(t^{FB}, p)] \\ &= \frac{F''(t^{FB})(pc_I + (1-p)c_E)}{B(t^{FB}, p)} + prF''(t^{FB}) \frac{A(t^{FB}, p)}{B(t^{FB}, p)^2} < 0 \end{aligned} \quad (74)$$

Since $V(t)$ is strictly concave at any interior critical point, $V(t)$ is strictly pseudoconcave (Avriel et al. (1988)), and $t^{FB}(p)$ is a unique, global maximum. It follows that $t^S(p) < t^{FB}(p)$. Therefore, the self routing scheme can never achieve first best. Since the agents' best response functions in dedicated and cross routing only depend on p and are different from the FOC of t^{FB} , first best cannot be implemented by these two schemes. ■

PROOF OF LEMMA 11. To show $t^S(p) < t^D(p)$, substituting $t^S(p)$ into the FOC of $t^D(p)$ yields that $\frac{w}{t^S(p)}[pt^S(p)f(t^S(p)) + p\bar{F}(t^S(p)) - 1] = \frac{w}{t^S(p)}[\frac{t^S(p)}{r} + p\bar{F}(t^S(p)) - 1] > 0$. We show $t^D(p) < t^C(p)$ by contradiction. Suppose $t^D(p) \geq t^C(p)$ and it follows from the FOC of $t^D(p)$ that $pt^C(p)f(t^C(p)) + p\bar{F}(t^C(p)) - 1 \geq 0$. Then,

$$\begin{aligned} & p[t^C(p)f(t^C(p)) + \bar{F}(t^C(p))][1 + (1 - \frac{r}{t^C(p)})\rho(t^C(p))] + \rho(t^C(p))^2 - 1 \\ &\geq 1 + (1 - \frac{r}{t^C(p)})\rho(t^C(p)) + \rho(t^C(p))^2 - 1 = \frac{\rho(t^C(p))}{t^C(p)}(t^C(p) + p\bar{F}(t^C(p))r - r) > 0, \end{aligned} \quad (75)$$

contradicting the FOC of $t^C(p)$. ■

PROOF OF LEMMA 12. Since the principal's objective function is monotonically decreasing in w in three routing schemes, all the IR constraints must bind except IR2 under dedicated routing. Therefore, $w^S(p) = \alpha(t^S(p) + p\bar{F}(t^S(p)r))$, $w^D(p) = \frac{\alpha t^D(p)}{1-p\bar{F}(t^D(p))}$, and $w^C(p) = \alpha(t^C(p) + p\bar{F}(t^C(p)r))$. The rest of the proof is omitted due to similarity to that of Lemma 6. ■

PROOF OF PROPOSITION 4. First we compare the principal's profit rate at any p ,

$$\begin{aligned}
& V^C(p) - V^S(p) \\
&= \frac{v - \bar{F}(t^C(p))(pc_I + (1-p)c_E) - C_A(p)}{t^C(p) + p\bar{F}(t^C(p))r} - \frac{v - \bar{F}(t^S(p))(pc_I + (1-p)c_E) - C_A(p)}{t^S(p) + p\bar{F}(t^S(p))r} \\
&= (t^C(p) - t^S(p)) \frac{(v - C_A(p))(pr \frac{F(t^C(p)) - F(t^S(p))}{t^C(p) - t^S(p)} - 1)}{(t^C(p) + p\bar{F}(t^C(p))r)(t^S(p) + p\bar{F}(t^S(p))r)} \\
&\quad + \frac{(pc_I + (1-p)c_E)(t^C\bar{F}(t^S(p)) - t^S\bar{F}(t^C(p)))}{(t^C(p) + p\bar{F}(t^C(p))r)(t^S(p) + p\bar{F}(t^S(p))r)} \tag{76}
\end{aligned}$$

(i) Since $t^C(p) > t^S(p)$, it follows that $\frac{F(t^C(p)) - F(t^S(p))}{t^C(p) - t^S(p)} < f(t^S(p)) = \frac{1}{pr}$ and $t^C\bar{F}(t^S(p)) - t^S\bar{F}(t^C(p)) > 0$. Therefore, $V^C(p) - V^S(p) > 0$ when $pc_I + (1-p)c_E$ is large enough or when $v - C_A(p)$ is small enough, i.e., if c_I or c_E are sufficiently large or if $C_A(\cdot)$ is sufficiently convex. Hence, $V^C = V^C(p^C) > V^C(p^I) > V^S(p^I) = V^S$. The first inequality follows from the optimality of V^C . (ii) $V^C(p) - V^S(p) < 0$ if v is large enough. Hence, $V^C = V^C(p^C) < V^S(p^C) < V^S(p^I) = V^S$. The last inequality follows from the optimality of V^S . ■

PROOF OF PROPOSITION 5. Since IR1 implies IR2, the principal's problems under dedicated and cross routing are identical, implying $C^D = C^C$. We will next compare C^C with C^S . First notice that the two optimization problems are identical except that cross routing has a less stringent IC constraint than self routing. This immediately implies that $C^C \leq C^S$. We next find conditions under which the inequality is strict. Let \hat{p} be the solution to the unconstrained optimization problem:

$$\min_{0 \leq p \leq 1, w \geq 0} g(t_H) + p(1 - \pi_H)g(r) + (1 - \pi_H)(pc_I + (1-p)c_E) + C_A(p) \tag{77}$$

The FOC is $C'_A(\hat{p}) = (1 - \pi_H)(c_E - c_I - g(r))$. Since the objective function is monotonically increasing in w under self routing, the IR constraint must bind. Therefore, finding p^S boils down to comparing \hat{p}^S with \hat{p} :

$$p^S = \begin{cases} \hat{p} & \text{if } \hat{p} > \bar{p}^S \\ \bar{p}^S & \text{if } \hat{p} \leq \bar{p}^S \end{cases} \tag{78}$$

Similarly if the IR constraint under cross routing binds, then

$$p^C = \begin{cases} \hat{p} & \text{if } \hat{p} > \bar{p}^C \\ \bar{p}^C & \text{if } \hat{p} \leq \bar{p}^C \end{cases} \tag{79}$$

We have shown that in the main text that $\bar{p}^C < \bar{p}^S$. Consider three possibilities: 1) If $\bar{p}^C < \bar{p}^S < \hat{p}$, then $p^C = p^S = \hat{p}$, which implies $C^C = C^S$; 2) If $\bar{p}^C < \hat{p} \leq \bar{p}^S$, then $p^C = \hat{p}$ and $p^S = \bar{p}^S$, which implies $C^C > C^S$; 3) If $\hat{p} \leq \bar{p}^C$, then $p^C = \bar{p}^C$ and $p^S = \bar{p}^S$, which implies $C^C > C^S$. From the first-order condition of \hat{p} , we conclude that case 1 will occur if c_E is sufficiently large and case 2 & 3 will occur if c_I is sufficiently large or $C_A(\cdot)$ is sufficiently convex.

Now we suppose the IR constraint under cross routing does not bind. Consider the Lagrangian of the constrained optimization problem:

$$\begin{aligned} L = & w + (1 - \pi_H)(pc_I + (1 - p)c_E) + C_A(p) - \phi[w - g(t_H) - p(1 - \pi_H)g(r)] \\ & - \sigma\left[p - \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)w}\right]. \end{aligned} \quad (80)$$

Taking the derivative yields

$$\frac{\partial L}{\partial p} = (1 - \pi_H)(c_I - c_E) + C'_A(p) + \phi(1 - \pi_H)g(r) - \sigma = 0, \quad (81)$$

$$\frac{\partial L}{\partial w} = 1 - \phi - \sigma \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)w^2} = 0. \quad (82)$$

Nonbinding IR constraint implies $\phi = 0$. From $\frac{\partial L}{\partial w} = 0$, we infer that $\sigma > 0$. Complimentary slackness implies that IC must bind i.e., $w = \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)p}$. Substituting into the objective function yields

$$\min_{0 \leq p \leq 1} \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)p} + (1 - \pi_H)(pc_I + (1 - p)c_E) + C_A(p). \quad (83)$$

Taking the derivative yields

$$- \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)p^2} + (1 - \pi_H)(c_I - c_E) + C'_A(p) = 0. \quad (84)$$

The SOC is

$$2 \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)p^3} + C''_A(p) > 0. \quad (85)$$

Substituting \hat{p} into the above FOC gives

$$- \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)\hat{p}^2} + (1 - \pi_H)(c_I - c_E) + C'_A(\hat{p}) = - \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)\hat{p}^2} - (1 - \pi_H)g(r) < 0, \quad (86)$$

which implies $\hat{p} < p^C$ due to strict convexity of the objective function. In addition, binding IC constraint implies $p^C = \bar{p}^C < \bar{p}^S$. Therefore, $p^S = \bar{p}^S$. Then,

$$\begin{aligned} C^C - C^S &= \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)\bar{p}^C} - g(t_H) - (1 - \pi_H) \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)} \\ &+ (1 - \pi_H)(\bar{p}^S - \bar{p}^C)(c_E - c_I) + C_A(\bar{p}^C) - C_A(\bar{p}^S). \end{aligned} \quad (87)$$

If c_I is sufficiently large or $C_A(\cdot)$ is sufficiently convex, then $C^C - C^S < 0$. Otherwise, $C^C = C^S$. ■