Chapter 2
The Basic Principal-Agent

In a basic principal-agent setting, the principal contracts an agent to perform a service function and the agent chooses the level of his capacity (his 'effort') in response to the contract offer and subsequently its effect on the principal's revenue stream. We assume that the principal's equipment unit generates revenue at an expected rate of $r > 0$ \$ per unit of uptime. The unit runs for a random period of time before failing, and remains in the failed state until it is repaired. To address the recurring maintenance and equipment failures the principal contracts an agent who subsequently installs a repair capacity and repairs the principal's equipment when it fails. The contract structure considered is rather simple: the principal proposes to pay the agent $w > 0$ \$ per unit of time during the duration of the contract but the agent pays the principal $p > 0$ \$ per unit of time during the unit's failure duration. The agent's capacity decision is unobservable by the principal. Each party is presumed to choose the values that maximize his/her utilities. We assume that the parties are rational and each knows that the other is rational, etc. till infinitum. It includes their individual computational ability to anticipate (compute) the other's best response to any offer. Therefore, with some abuse of timing we presume that both, the contract offer and the service capacity decision, occur at the same time with full knowledge of the two parties.

In general, if the agent's action is observable and contractible, then the principal would contract directly on agent's service capacity that maximizes the principal's profit leaving zero surplus to the agent – enough to ensure agent's participation. Such a scenario is referred to as the *first-best solution* (Hölmstrom 1979). If the agent's action is unobservable and therefore uncontractible, then the agent's response may deviate from the one prescribed by the principal in the first-best solution, and the principal risks realizing lower profits. The likelihood and the degree of agent's deviation from the desired action is referred to as *moral hazard* (Luenberger 1995). When moral hazard is present the principal uses the available

© Springer International Publishing Switzerland 2016

S. Zeng, M. Dror, *Formulating Principal-Agent Service Contracts for a Revenue Generating Unit*, SpringerBriefs in Operations Management,
DOI 10.1007/978-3-319-18672-6_2

information about the agent's action to alleviate the moral hazard (Hölmstrom 1979) and proposes a contract with incentives that aim the agent to maximize her profit.

Principal's main information about the agent's capacity is deduced from her revenue stream. The revenue consequences of agent's action are referred to as the service performance characteristics, and quantified service performance metrics are referred to as performance measures. The contracts that use performance measures are called performance based contracts. By offering an agent performance based contract, the principal transfers part of her risk regarding revenue to the agent's revenue risk, thus providing incentives for the agent to choose the action desired by the principal. If the performance measure is positively correlated with principal's revenue, a rate of award for each unit of the performance measure, known as the piece rate b, is specified in the contract. If the performance measure is negatively correlated with principal's revenue, a penalty rate for each unit of the performance measure, denoted by p, is specified in the contract.

Under performance based contracts, the agent maximizes his utility based on the scheme proposed by the principal, and the principal maximizes her profit while anticipating the agent's optimizing decision. This scenario is referred to as the *second-best solution* (Hölmstrom 1979). Given a compensation scheme, if the agent's utility is globally concave, the second-best solution can be derived using first order condition of the agent's utility, referred to as the *first-order approach*. If the agent's utility is not globally concave, the first-order approach is generally invalid and alternative approaches have to be used such as converting the agent's utility optimization problem into a convex programming problem (Grossman and Hart 1983).

In our case short unit's downtimes (relative to uptimes) imply a higher revenue for the principal, thus the downtimes and their frequency infer the agent's service performance. The service capacity can only be inferred to by the nature of downtimes, which are unobservable before signing the contract. Therefore moral hazard is of concern with performance based contracts. The performance measure adopted here is based on the unit's downtimes. The downtimes are negatively correlated with principal's revenue, and the agent is charged a penalty p \$ for each unit (seconds, minutes, hours or days) of the performance measure.

In Kim et al. (2010) the profit function of the principal and the utility function of the agent are based on three assumptions. First, the unit is mission-critical and the principal owns one unit. Second, the unit is highly reliable such that the service times are relatively short as compared to the uptimes. Third, the service times are independently and identically distributed, and the distribution has no upper bound on the realization of the service times. This model has two pitfalls: (i) Kim et al. (2010) assume the failures as a Poisson arrival process independent of the service times. It allows for a new failure to occur while the unit is still in a failed state, contradicting that no new failure can occur when in a failed state. (ii) The profit/utility functions describe the total profit/utility during a single contract period assumed finite and normalized to 1. Although the contract period is finite, it contradicts their assumption about the service time distribution with no upper bound on duration of the service time.

Table 2.1 The variables of the model

Variable	Description	Type
η	Agent's risk attitude	Exogenous
r	Unit's revenue rate	Exogenous
λ	Unit's failure rate	Exogenous
c	Marginal rate of capacity cost	Exogenous
w	Agent's compensation rate	Determined by the principal
p	Agent's penalty rate	Determined by the principal
μ	Service capacity	Determined by the agent

To repeat, the failure rate of the equipment unit is a constant λ, the repair time is exponential with a constant repair rate μ (the service capacity is the repair rate), yielding a less general model than Kim et al. (2010). Furthermore, we do not restrict the contract to a period of time, rather, the contract can be dynamic and can be offered and accepted/rejected continuously in time.

The unit's failure rate $\lambda > 0$, the principal's expected revenue rate $r > 0$, and the marginal capacity cost $c > 0$, are exogenous variables. The payment rate w and the penalty rate p are determined by the principal, whereas the service capacity $\mu \geq 0$ is determined by the agent. We denote an exogenous scalar parameter η as preference and intensity indicator for agent's risk attitude: $\eta = 0$ for risk-neutral, $\eta > 0$ for risk-averse, and $\eta < 0$ for risk-seeking.

The seven variables that appear in our model are listed in Table 2.1.

Two performance measures are considered in Kim et al. (2010). The first one is *cumulative downtime* – the sum of downtimes during a finite contract period. The second one is the *average downtime*, which uses the sample average of downtimes during a finite contract period as the performance measure. The two measures provide different incentives for the agent's capacity decisions. In essence, the agent's optimal service capacity behaves non-monotonically with the failure rate when using average downtime, while it is monotonically increasing when using cumulative downtime. This is because average downtime reflects the risk differently compared to cumulative downtime. When the failure rate is higher, the expected number of failures is higher during the finite contract period. For a higher number of failures and the same service capacity, average downtime dilutes the agent's risk by a factor proportional to the square of the number of failures as compared to cumulative downtime, thus provides an incentive for the agent to choose a lower service capacity, leading to reduced service performance. We adopt the steady state probability of the failed state as the sole performance measure, which is the equivalence of cumulative downtime in our undetermined time horizon setting.

The literature on principal-agent setting is extensive in economics since the topic is fundamental to the economic analysis of firms' interdependence via contractual agreements that impact their output. We do not survey here the principal-agent literature. This has been done very well by numerous authors. A partial list includes Ross (1973), Hölmstrom (1979), Stiglitz (1974, 1979), Myerson (1983), Hölmstrom

and Milgrom (1987), Fudenberg and Tirole (1990), Maskin and Tirole (1990, 1992), and Bolton and Dewatripont (2005). For analytic and numerical solutions to principal-agent problems see Grossman and Hart (1983) and Guesnerie and Laffont (1984).

2.1 Contractual Relationship Between a Principal and an Agent

When an agent contracts a single principal, the agent is always available when the unit fails, therefore the unit's downtimes are the same as the service times. To mitigate the pitfalls in Kim et al. (2010) we recast this system a Markov process. The state of the Markov process is defined as the state of the principal's unit: in state 0 when the unit is operational, and in state 1 when the unit is not operational. We assume that the uptimes of the unit are independently and identically distributed following an exponential distribution that is governed by the unit's failure rate, and the service times of the unit are independently and identically distributed, following an exponential distribution governed by the agent's service capacity. For a risk-neutral agent we propose an objective function that describes his expected utility rate for each unit of time in an infinite time contract assuming the Markov process is in steady state. Similarly we propose an objective function that describes a risk-neutral principal's expected profit rate. Both the principal's and the agent's objective functions depend on the compensation rate $w > 0$ paid by the principal to the agent and the penalty rate $p > 0$ charged by the principal for each unit of downtime. Furthermore, the principal's expected profit rate also depends on the revenue rate $r > 0$, and the agent's expected utility rate also depends on the marginal cost $c > 0$ of the service capacity for each unit of time. In our principal-agent contractual relationship, the principal controls w and p, and the agent controls μ, therefore we call vector $((w, p), \mu)$ a *strategy*. The c is exogenously determined by the market and in this paper it is normalized as a monetary unit $\Rightarrow c \equiv 1$. Observation 3.1 (below) points out that a contract with compensation rate w paid only for each unit of uptime and penalty rate charged for each unit of downtime is equivalent to our setting of principal-agent contract.

Notation: Denote the principal's expected profit rate by $\Pi_P(w, p; \mu)$ and the agent's expected utility rate by $u_A(\mu; w, p)$, omitting the exogenous parameters.

When the agent does not accept the contract offer he commits no service capacity and receives no compensation. $\underline{u_A}(\mu = 0) = 0$ is referred to as *the agent's reservation utility rate*. An agent accepts the contract only if his expected utility rate is greater than or equal to his reservation utility rate, referred to as the individual rationality (IR) constraints. When the principal does not contract an agent for the repair service, then since an equipment failure will occur after some finite time with probability 1, therefore in the long run the principal's expected profit rate equals zero, which is referred to as *the principal's reservation profit rate* ($\underline{\Pi_P} = 0$).

Individual rationality principal dictates that the principal offers a contract only if her expected profit rate is strictly greater than her reservation profit rate.

When a principal-agent contract exists, the agent's average utility over a finite period of time converges to his expected utility rate as the period approaches infinity. However it is still probable that the agent receives negative revenue stream over some finite period of time, such that his cumulative revenue (utility) drops below a certain threshold and triggers bankruptcy preference claim against the agent. In our paper, we presume that the likelihood of such bankruptcy condition to occur is negligible.

The above principal-agent problem is characterized by expression of the principal's and agent's expected profit/utility rates and the values of the exogenous parameters. Denote a principal-agent problem by $\mathfrak{P}(\Pi_P, u_A, \eta, \lambda, r)$ or for short $\mathfrak{P}$.

Definition 2.1 (*Strategy Set*). *The* ***strategy set*** *of a principal-agent problem* $\mathfrak{P}$ *is defined as a vector* $\mathfrak{S}(\mathfrak{P}) \equiv \{((w,p),\mu)\,|w > 0, p > 0, \mu \geq 0\}$.

Definition 2.2 (*Weak Domination*). *Consider two strategies* $((w,p),\mu), ((w',p'),\mu') \in \mathfrak{S}(\mathfrak{P})$. $((w,p),\mu)$ *is said to* ***weakly dominates*** $((w',p'),\mu')$, *denoted by* $((w,p),\mu) \succeq ((w',p'),\mu')$, *if the two strategies result in* $\Pi_P(w,p;\mu) \geq \Pi_P(w',p';\mu')$ *and* $u_A(\mu;w,p) \geq u_A(\mu';w',p')$ *with at least one strict inequality.*

Definition 2.3 (*Set of Admissible Solutions*). *The* ***set of admissible solutions*** (*also known as the* ***set of Pareto optimal solutions***) *for the principal-agent problem* $\mathfrak{P}$ *is the set* $\mathfrak{s}(\mathfrak{P})$ *of all strategies* $((w,p),\mu) \in \mathfrak{S}(\mathfrak{P})$ *for which:*

(a) $\nexists\ ((w',p'),\mu') \in \mathfrak{S}(\mathfrak{P})$ *such that* $((w',p'),\mu') \succeq ((w,p),\mu)$ – *there is no other strategy that weakly dominates* $((w,p),\mu)$.
(b) $\Pi_P(w,p;\mu) > \underline{\Pi_P}$ *and* $u_A(\mu;w,p) \geq \underline{u_A}$.

Pareto optimality implies that the principal cannot increase her expected profit rate without lowering the agent's expected utility rate and vice versa (Luenberger 1995), and it has been proven that generally both the principal and the agent achieve Pareto optimality as a subset of the second-best solutions (Ross 1973). Since the agent's IR is always binding, condition (a) in Definition 2.3 guarantees that all admissible solutions are Pareto optimal. We require that all the solutions proposed in this paper be *Admissible Solutions*.

This paper is organized as follows. In Chap. 3, we present the basic model with a risk-neutral principal and a risk-neutral agent, and we describe the exogenous conditions that guarantee the existence of a contract and the optimal contract terms. In Chap. 4 we analyze risk-averse agent. Chapter 5 is dedicated to the analysis of a risk-seeking agent. In Chap. 6 we summarize our findings and conclusions. Notation is introduced as needed.