We consider a monopolist expert offering a service with a ‘credence’ characteristic. A credence service is one where the customer cannot verify, even after a purchase, whether the amount of prescribed service was appropriate or not; examples include legal, medical or consultancy services and car repair. This creates an incentive for the expert to ‘induce service’, that is, to provide unnecessary services that add no value to the customer, but that allow the expert to increase his revenues. We focus on the impact of an operations phenomenon on service inducement - workload dynamics due to the stochasticity of inter-arrival and service times. To this end, we model the expert’s service operation as a single-server queue. The expert determines the service price within a fixed and variable fee structure and determines the service inducement strategy. We characterize the expert’s combined optimal price structure and service inducement strategy as a function of service capacity, market potential, inducement opportunity, value of service and waiting cost. We find that service inducement is a means to dynamically skim customer surplus with state-independent prices and provision of slower service to customers that arrive when the expert is idle. We conclude with design implications of our results in limiting service inducement.
1 Introduction

In many service contexts, customers do not know the appropriate level of service required for a complex product or operation. They rely on the advice of an ‘expert’ who typically also provides the subsequent service. Furthermore, it is difficult for the customer to verify whether the provided service was appropriate, even after the service is performed. Darby and Karni (1973) coined the name ‘credence good’ for a good whose quality cannot costlessly be ascertained by the customer even after purchasing it. This is in contrast to an ‘experience good’ for which usage reveals quality. Examples of credence goods are medical, legal and repair services. In such a setting, if selling more services than what is really required allows the expert to make a higher profit, a moral hazard problem is created: The expert has an incentive to perform unnecessary service. We refer to this phenomenon as ‘service inducement.’

One key element in the expert’s incentive to induce service is the fee structure. A fee that is proportional to the level of service provided makes it feasible for the expert to induce service profitably. In contrast, a fixed fee makes it unprofitable for the expert to induce service. This is because with a fixed fee, the expert is not compensated for any additional - albeit unnecessary - service.

In their seminal paper, Darby and Karni qualitatively discussed, but did not analyze, another key element that impacts the expert’s incentive to induce service, the expert’s dynamically changing workload level: At low workloads, the expert would benefit more from inducing service than at high workloads since service inducement increases the congestion of the system. Since in many service settings customer inter-arrival times and service times are stochastic, the expert’s workload does fluctuate over time. Thus, both price structure and workload dynamics impact the expert’s choice of service inducement strategy in equilibrium. In this paper, we develop a model that allows us to better understand how these two factors impact the expert-customer interaction.

To this end, we consider a monopolist expert selling a single service. Workload dynamics are modeled using a single-server queue with a Poisson arrival process of potential customers, and independently and identically distributed value-adding service times. Customers are homogeneous in that they place the same value on the service and have the same waiting cost per unit time. The service price consists of a fixed fee and/or a variable component that is proportional to the total service time. The expert determines the fixed
and variable fees, and chooses a service inducement policy. Customers observe the system state upon arrival and decide whether to purchase the service (i.e. join the queue or not) based on the net utility that they expect from obtaining that service.

Elements of our stylized model can be found in different service industries. An example is car repair: “One of the largest sources of complaints received by consumer protection agencies involves [...] repair. [...] Consumers are exposed to all manners of rip-offs, scams and blatant overcharges. [...] The types of scams that can befall you [...] include adding repairs that are really not needed [...]” (http://www.carclicks.com).

Another example is legal advice. Many legal services are billed proportional to the time that an attorney puts into the service, sometimes to the tenth of an hour (Ross 1996). Drawing on his surveys, the experiences of legal audit firms, and anecdotes, Ross concludes that the incentive to over-billing is widespread among attorneys. Much of the ‘padding’ of hours is impossible to detect and “can escape the attention of even the most dedicated sleuth” (p. 23). Ross also mentions that the incentives to over-bill depend on the workload level. In particular, padding occurs at times that lawyers are ‘not busy’ since they have the time to do unnecessary tasks (p. 36-37).

A recent article (Economist 2006) discussing credence goods states “New York taxis charge $2.50 the moment you sit in them, and another $2 for every mile covered.... When they have no trouble finding fares, taxi-drivers have no reason to take you the long way round. If they were not serving you they could be making as much money, or more, serving the next person. In quiet periods, however, the opportunity cost of ‘overtreating’ clueless passengers falls, and the rewards rise.” This is another example of how incentives to induce service depend on fluctuating workload levels.

Finally, there is an ongoing debate in the health care literature about the existence of physician-induced demand. In a recent empirical study, Delattre and Dormont (2003) show evidence of physician-induced demand in France. They find that the number of consultations per doctor only slightly decreases with an increase in the physician/population ratio. In addition, physicians counterbalance the fall in the number of customers by an increase in the volume of care delivered in each encounter. In other words, utilization impacts physician-induced demand.

Motivated by these examples, we address the following questions: How do characteristics of the environ-
ment, in particular, service capacity, inducement opportunity, market size, value of service and waiting cost
determine the expert’s incentive to induce service? What is the structure of the service inducement policy?
How does a credence service provider set the service fee (fixed and/or variable) differently from a provider
of a non-credence service?

Our analysis is not a normative one: We do not wish to provide advice about when it is optimal to induce
service; our goal is to generate insights for managers of service systems about drivers that impact choices
of service providers. Existing research on credence goods focuses on the impact of capacity, reputation and
competition on the existence of service inducement. Our research complements this literature by focusing on
the role that workload dynamics - an operations phenomenon - plays in the provision of service inducement.
We also complement the research that uses queuing models to analyze service systems, which to the best of
our knowledge, ignores the credence character of certain services.

We obtain several insights: According to existing research (e.g. Dulleck and Kerschbamer 2006), service
inducement can never be optimal when the customer base is homogenous with respect to their willingness
to pay for the expert’s service. Even though in our model, the customer base is homogenous, differing
congestion levels upon arrival creates effective customer heterogeneity. Two prices (a fixed fee and a billing
rate) and the operational flexibility to provide slower service to customers arriving when the expert is idle,
to which we refer as ‘skimming,’ allow the expert to appropriate a larger share of the total welfare generated
than with honest (i.e. fast) service. However, skimming reduces the total welfare generated compared to an
expert who always provides honest service. The profitability of service inducement depends on which effect
dominates: The level of total social welfare versus how efficiently the expert appropriates it as his own profit.
We uncover fundamental determinants for service inducement to be profitable: (1) waiting costs should not
be too high, (2) the arrival rate of potential customers for the expert should neither be too high nor too low.
These conditions strongly depend on the service inducement potential, which is a measure of how much the
expert can slow down the service: When the service inducement potential is high, service inducement may
not be profitable. When the potential decreases, service inducement becomes more profitable over a wider
range of parameter settings.

The remainder of the paper is structured as follows: §2 puts our work in the context of the existing
literature and highlights our contributions. Our modeling assumptions are described in §3. We characterize the expert’s optimal prices and profits without and with inducement in §4. In §5, we build on the analysis in §4 to determine when service inducement is more profitable for the expert. §6 discusses the main insights from our analysis and the design implications of our results.

2 Related Literature

The model we develop for our analysis draws on the queueing literature that takes into account the strategic interaction between the server and the customer. Such a strategic interaction in a queueing context was first studied by Naor (1969). This paper and the subsequent literature (for an excellent overview, see Hassin and Haviv 2003) study the impact of congestion on the customers’ and service provider’s decisions. In particular, Hassin (1986) characterizes the equilibrium fixed fee in a single-server observable queue with a homogeneous customer base, Poisson arrivals and exponential service times. Van Mieghem and Lariviere (2004) examine how the negative externalities due to congestion induce strategic customer behavior, without analyzing the levers available to a service provider in influencing customer behavior. To the best of our knowledge, our paper is the first to model the ‘credence good’ characteristic of services in this literature.

Early papers on credence goods (Darby and Karni 1973, Glazer and Hassin 1983, Pitchik and Schotter 1987a,b) develop simple models and identify the existence of service inducement. Recently, models allowing endogenous pricing in monopoly (Emons 2001, Fong 2005) and competitive settings (Wolinsky 1995, Emons 1997, Richardson 1999, Pesendorfer and Wolinsky 2003, Alger and Salanie 2006) have been developed. Of these papers, only Emons models the expert as being capacity-constrained, but uses a simple deterministic model. Our contribution to this literature is to develop a richer model of a capacitated monopoly service system that explores the role of workload dynamics; this issue has been qualitatively discussed but not analyzed in Darby and Karni’s seminal paper. Below, we position our work with respect to papers analyzing the monopoly case and discuss our contributions in more detail.

Fong (2005) shows that in an un-capacitated system with homogeneous customers, charging a fixed fee regardless of service type is optimal and eliminates service inducement. Emons (2001) considers a capacity-constrained monopolist serving a homogeneous customer base who determines the capacity level and prices
of diagnosis and repair. The time required to serve each customer honestly is identical and deterministic. Emons finds that charging a fixed fee or having a capacity level exactly equal to that required to serve the whole market honestly (100% utilization) are sufficient to signal credibility. Note that with stochastic inter-arrival and service times, 100% capacity utilization is not viable, so the expert can only use pricing as a mechanism to signal credible service, which is confirmed by our analysis. Like Fong who studied the un-capacitated case, Emons finds that serving the market honestly is optimal for the capacitated expert. In contrast, we demonstrate that even with a homogeneous customer base, workload dynamics results in the emergence of service inducement under some conditions.

The difference can be explained based on Dulleck and Kerschbamer (2006), who develop a simple model unifying the literature and delineating drivers impacting the existence or non-existence of service inducement in equilibrium. They identify customer homogeneity (an assumption made by both Fong and Emons) as one of the necessary conditions to eliminate service inducement. The logic is as follows: With a homogeneous customer base, a single price that leaves each customer indifferent between purchasing service or not exists. This price extracts all consumer surplus and maximizes expert profit. Since unnecessary service inducement only “destroys” consumer surplus, the most profitable strategy for the expert is to not induce service. Only with a heterogeneous customer base may the expert find it profitable to induce service. In particular, since the expert is not able to capture all surplus using a single price, he may find it optimal to selectively induce service to some customer types. When workload dynamics are taken into account, a customer base that is homogeneous with respect to service value and waiting cost becomes effectively heterogeneous upon arrival: Customers arriving at different times face different waiting times, which yields different levels of expected net utility from service. Our main contribution is to show that under some conditions, this heterogeneity makes service inducement more profitable for the expert, and to identify these conditions.

3 The model

In this section, we outline our assumptions regarding the characteristics and strategy of the customer base, the service characteristics, and the pricing and service inducement strategies of the expert. We end with the specification of the expert-customer game.
The customers. We consider an infinite, but countable number of risk neutral, short-lived players, referred to as ‘the customers’. Customers are labeled by $k \in N$. We assume that customers only partially observe the system: An arriving customer observes state $i$ upon arrival, indicating whether the expert is idle ($i = 0$); or not ($i = 1$). This modeling choice allows us to capture state-dependent customer decisions and is simple enough that we can analyze the service inducement phenomenon in closed form.

The customer decides whether to enter service or not based on his expected net utility from service. Let $S^k_i \in \{\text{join}, \text{balk}\}$ for $i \in \{0, 1\}$ be customer $k$’s action set when the state upon arrival is $i$. Customer $k$’s strategy is characterized by $\beta^k = (\beta^k_0, \beta^k_1) \in [0, 1]^2$, where $\beta^k_i = \Pr(S^k_i = \text{join})$ is the probability that customer $k$ enters if the system upon arrival is in state $i$. The customer base is homogeneous: All customers place value $V$ on the service, and incur a cost of $c$ per unit time spent waiting in queue or in service. We assume an additive utility structure where a customer’s net utility is $V$ minus the expected waiting cost minus the expected payment. Customers are risk neutral. If a customer decides not to enter, he obtains 0 utility.

Customers arrive according to a Poisson process of rate $\Lambda$. We call $\Lambda$ the ‘market potential’; this is the maximum rate at which customers may enter service. The effective arrival rate at the expert may be lower than $\Lambda$ because not all customers choose to enter the service system; this rate is determined by the customer strategy $\beta^k$. Since our focus is the impact of workload dynamics on the expert’s incentives to induce service, we do not incorporate the expert’s concern for repeat services: Each customer in the Poisson stream represents a new customer who does not have a history of transactions with the expert. For example, Callahan (2004) notes that due to the anonymity of corporate law, “... there is little loyalty between law firms and clients” (p. 35). The Poisson arrival stream generating new customers is particularly appropriate under such circumstances.

The expert and service characteristics. We consider one infinitely long lived player, called ‘the expert’. We assume that the expert has a monopoly position in the market. The expert decides the prices and the service inducement policy to maximize his expected steady-state profit rate. We assume that the expert posts a non-negative price vector $R = (R, r) \in \mathbb{R}^2_+$, where $R$ is a fixed fee charged to each customer that obtains the service and $r$ is the variable rate per unit of service time, also called the ‘billing rate’. 
We assume that there are two procedures that deliver the value $V$ to the customer. The service times of both procedures are exponentially distributed. One procedure has an expected service time of $\frac{1}{\mu}$. The other one has an expected service time of $\frac{1}{\mu'}$, where $\mu' > \mu$. Depending on the state of the expert (idle or not) upon arrival of the customer, the expert decides which procedure he will use for that customer. The expert’s service policy is denoted by $\mu = (\mu_0, \mu_1) \in \{\mu, \mu'\} \times \{\mu, \mu'\}$, where $\mu_i$ is the service rate applied to a customer arriving in state $i \in \{0, 1\}$. Based on the procedure and the price structure, $R$, the customer decides whether to buy the service, s/he joins a queue that the expert serves the customers in a first-come first-serve manner. This way of modeling state-dependent service rates is similar to Harris (1966). In our model, by selecting $\mu$ instead of $\mu'$, the expert extends the service time without creating more value for the customer. We refer to selecting $\mu$ as ‘service inducement.’

We refer to $\mu$ as the base service rate. The ratio $\alpha = \frac{\mu}{\mu'}$ represents the ‘inducement opportunity’; a ratio close to zero means that with the slow rate, the expert adds a significant amount of unnecessary service time. A ratio close to one means that the expert adds very little unnecessary service time.

Service cost is assumed to be 0 per unit time. We assume that the time spent by the expert on service is verifiable, either by the customer or by some agency. This means that the expert cannot claim to have done work without actually doing it. This assumption ensures that service inducement has an implicit ‘cost’ to the expert - it uses up limited capacity. We assume that incomplete servicing can be detected by the customer and that an institution exists where the customer can hold the expert liable for incomplete servicing. Therefore, the work content of a customer must be completed; e.g. the car has to be in working condition. This assumption reflects the fact that symptoms of the problem the customer wanted solved will persist if the expert does not provide the appropriate service. On the other hand, if service inducement occurs the expert is not penalized. This assumption reflects the fact that it is typically very difficult to show that no unnecessary service has been done (e.g. Ross 1996); the customer only observes that the problem has been solved.

Assumption 1 $V - 2c > 0$

This assumption ensures that the value of service is sufficiently high with respect to the waiting costs and allows us to eliminate uninteresting cases. The assumption will be discussed in section 4.2.
Specification of the game. We consider a two-stage game. In the first stage, the expert determines \( R = (R, r) \) and service policy \( \mu \); in the second stage, customers \( k = 1, 2, \ldots \) set \( \beta^k \). We assume that the price structure and all other parameters \((V, c, \mu, \Lambda)\) are common knowledge.

Formal description of the game. Formally, the expert’s strategy is composed of \( R \in \mathbb{R}_+^2 \) and a pair \( \mu \in \{\mu, \overline{\mu}\} \times \{\mu, \overline{\mu}\} \). Customer \( k \)'s strategy is characterized by a mapping from \((R, \mu)\) to \( \beta^k \in [0, 1] \times [0, 1] \).

Since all customers are homogeneous ex ante, we analyze symmetric customer equilibria \( \beta^k = \beta \) for all \( k \), as in the literature.

The Second Stage. For a given \( R \) and \( \mu \), the joining strategy of all other customers, \( \beta \), uniquely determines the expected utility that a randomly arriving customer observing state \( i \) obtains from joining, which we denote by \( V_i (\beta; \mu, R) \). The customer’s best response set in each state \( i \in \{0, 1\} \) is either to enter \((\{1\})\), not to enter \((\{0\})\) or to randomize between two actions with any probability \(([0, 1])\) when \( V_i \) is positive, negative or zero:

\[
BR_i (\beta; R, \mu) = \begin{cases} 
\{1\} & \text{if } V_i (\beta; \mu, R) \geq 0, i \in \{0, 1\} \\
[0, 1] & \text{if } V_i (\beta; \mu, R) < 0, i \in \{0, 1\} \\
\{0\} & \text{otherwise}
\end{cases}
\]

Symmetric customer equilibria for a given \((R, \mu)\) satisfy the following condition:

\[
\beta^*_i \in BR_i (\beta^*; R, \mu), i \in \{0, 1\}.
\]

In words, when fixing the strategy of all other customers to \( \beta^* \), if the best reaction set of a randomly arriving customer (which is jointly determined by \( \beta^* \) and the expert’s best reaction to \( \beta^* \)) contains \( \beta^* \), then, \( \beta^* \) is an equilibrium. If there are multiple strategies that satisfy these conditions, we select the strategy that yields the highest customer surplus. If all strategies yield the same surplus, the strategy with the highest joining rate is selected.

The First Stage. Let \( \Pi (\beta; R, \mu) \) be the expert’s profit rate when the customer strategy is \( \beta \) and the expert’s prices and service policy is \((R, \mu)\). The expert’s problem is to set the pricing and service strategy that maximizes his expected profit rate; \( \max_{\mu \in \{\mu, \sigma\} \times \{\mu, \sigma\}} \Pi (\beta^* (R, \mu); R, \mu) \). In our analysis, we first find \( \hat{\Pi} (\mu) = \max_{R \in \mathbb{R}_+^2} \Pi (\beta^* (R, \mu); R, \mu) \) which is the expert’s optimal expected profit rate for a given service strategy, \( \mu \), and then, we find \( \max_{\mu \in \{\mu, \sigma\} \times \{\mu, \sigma\}} \hat{\Pi} (\mu) \).
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Lambda$</td>
<td>arrival rate (market potential)</td>
</tr>
<tr>
<td>$\bar{\mu}$</td>
<td>base (honest) service rate</td>
</tr>
<tr>
<td>$\mu_s$</td>
<td>slow service rate</td>
</tr>
<tr>
<td>$V$</td>
<td>utility of customer from service</td>
</tr>
<tr>
<td>$c$</td>
<td>customer waiting cost per unit time</td>
</tr>
<tr>
<td>$k = 1, 2, \ldots$</td>
<td>customer index</td>
</tr>
<tr>
<td>$n = 0, 1, 2, \ldots$</td>
<td>system state</td>
</tr>
<tr>
<td>$i \in {0, 1}$</td>
<td>state observed by arriving customer</td>
</tr>
</tbody>
</table>

**Derived parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho \equiv \frac{\Lambda}{\bar{\mu}}$</td>
<td>base load</td>
</tr>
<tr>
<td>$\alpha \equiv \frac{\mu}{\bar{\mu}}$</td>
<td>inducement potential</td>
</tr>
<tr>
<td>$c' \equiv \frac{c}{V\bar{\mu}}$</td>
<td>normalized waiting costs</td>
</tr>
</tbody>
</table>

**Variables**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu = (\mu_0, \mu_1)$</td>
<td>service policy of the expert when $i \in {0, 1}$</td>
</tr>
<tr>
<td>$\mu_h = (\bar{\mu}, \bar{\mu})$</td>
<td>honest policy</td>
</tr>
<tr>
<td>$\mu_s = (\mu, \bar{\mu})$</td>
<td>skimming policy</td>
</tr>
<tr>
<td>$\beta = (\beta_0, \beta_1)$</td>
<td>general joining policy of the customer</td>
</tr>
<tr>
<td>$R = (R, r)$</td>
<td>price vector consisting of a fixed fee and a billing rate</td>
</tr>
<tr>
<td>$V_i (\beta, \mu, R)$</td>
<td>expected net utility of customer arriving in state $i$</td>
</tr>
<tr>
<td>$\Pi (\beta, \mu, R)$</td>
<td>expert’s steady-state profit rate for given $\beta, \mu, R$</td>
</tr>
<tr>
<td>$\hat{\Pi} (\mu)$</td>
<td>expert’s optimal profit rate for given $\mu$</td>
</tr>
</tbody>
</table>

Table 1: Summary and explanation of the main symbols used in the paper.

With this game, we capture two effects that are of interest in our context: First, the congestion effect where for a given price vector, the utility of a randomly arriving customer depends on the other customers’
decisions and the service policy. Second, the service inducement effect where for a given price vector, the expert’s service inducement policy depends on the customer’s decisions. Table 1 summarizes the main notation used in this paper.

4 Analysis

In this section, we analyze the subgame perfect, symmetric equilibria of the game that we described in the previous section. Section 4.1 derives some useful preliminary results about the waiting times and the utilization rate of the expert. Section 4.2 determines the customer joining strategy for a given price structure and service policy. Section 4.3 determines the optimal prices and profits for a given service strategy. In section 5, we discuss the optimal service strategy.

4.1 Preliminary Results

We start by defining the expected utility that a customer arriving when the expert is idle \( V_0 \) or busy \( V_1 \) would obtain if he joins the queue assuming that the expert’s policy is \((R, \mu)\) and all other customers follow policy \( \beta \):

\[
\begin{align*}
V_0 (\beta, R, \mu) &= v_0 (R, \mu) \\
V_1 (\beta, R, \mu) &= v_1 (R, \mu) - cW (\beta, \mu) .
\end{align*}
\]

Here,

\[
v_i (R, \mu) = V - c \frac{1}{\mu_i} - \left(R + \frac{r}{\mu_i}\right)
\]

is the net utility of the service, \( V \), minus the expected waiting cost in service minus the total expected payment when a customer arrives in state \( i \), \( v_i (R, \mu) \); and \( W \) is the expected waiting time in the queue of the customer that arrives when the expert is busy. When the customer arrives when the expert is idle \((i = 0)\), he does not incur any queuing costs upon joining.

The expert’s profit rate is composed of the fixed fees that he collects from customers arriving when the expert is idle (with rate \( \Lambda_0 \)) or busy (with rate \( \Lambda_1 \)) and the billing rate that he collects when he is busy:

\[
\Pi (\beta, R, \mu) = R (\Lambda_0 (\beta, \mu) + \Lambda_1 (\beta, \mu)) + r (1 - \pi_0 (\beta, \mu)) ,
\]
where \( \pi_0(\beta, \mu) \) is the steady-state probability that the expert is idle, given \( \beta \) and \( \mu \), and

\[
\Lambda_0(\beta, \mu) = \beta_0 \pi_0(\beta, \mu) \Lambda \quad \text{and} \quad \Lambda_1(\beta, \mu) = \beta_1 (1 - \pi_0(\beta, \mu)) \Lambda
\]

are the customer arrival rates when the expert is idle and busy, respectively. The social welfare generated is

\[
SW(\beta, \mu) = V(\Lambda_0(\beta, \mu) + \Lambda_1(\beta, \mu)) - cQ(\beta, \mu) \tag{2}
\]

where \( Q \) is the expected number of customers in the system; the social welfare is the expected value generated for joining customers minus the expected cost of the time spent in the system. Due to the memoryless property of the service and arrival times, the number of customers in the system can be modeled as a Markov process, whose steady-state probabilities can easily be found in closed form. Proposition 1 derives \( W, \pi_0 \) and \( Q \) that determine \( V_0, V_1, \Pi \) and \( SW \).

**Proposition 1** The expected queuing time of a customer joining the queue when the expert is busy is

\[
W(\beta, \mu) = \frac{1}{\mu_1} + \frac{1}{\mu_0} - \frac{1}{\mu_1}
\]

The probability that the expert is idle is

\[
\pi_0(\beta, \mu) = \frac{1 - \beta_1 \Lambda}{\mu_1} + \frac{\beta_0 \Lambda}{\mu_0}
\]

and the expected number of customers in the system is

\[
Q(\beta, \mu) = \frac{\beta_0 \Lambda}{\mu_0} - \frac{2}{\mu_0} - \frac{\beta_1 \Lambda}{\mu_1} \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right)
\]

for any \( \beta_1 \Lambda < \mu_1 \).

We use these expressions in the next section to determine the equilibrium customer joining strategy for a given price structure and service strategy \((R, \mu)\).

### 4.2 Customer joining strategy for a given price structure and service policy

In this subsection, we derive the customer joining strategy \( \beta^* \) for a given price structure, \( R \), and service policy, \( \mu \). As \( V_0(\beta, R, \mu) \) is independent of \( \beta \), the equilibrium conditions for \( \beta^*_0 \), the equilibrium joining decision when the expert is idle, can be determined independently of the condition for \( \beta^*_1 \). Lemma 2 characterizes \( \beta^*_0 \).

11
Lemma 2 For a given \((R, \mu)\), \(\beta_0^* (R, \mu)\) is determined as follows:

\[
\beta_0^* (R, \mu) = \begin{cases} 
1 & \text{if } v_0 (R, \mu) > 0 \\
[0, 1] & \text{if } v_0 (R, \mu) = 0 \\
0 & \text{if } v_0 (R, \mu) < 0.
\end{cases}
\]

The equilibrium condition for \(\beta_0^*\) is intuitive: When the expected utility from joining is positive (negative), all (no) customers join. Otherwise, when \(v_0 (R, \mu) = 0\), customers are indifferent and can choose to randomize with any probability of joining in \([0, 1]\). There are thus multiple equilibria possible for a given \((R, \mu)\) for which the expected customer utility is the same (zero). In the remainder of the paper, when \(v_0 (R, \mu) = 0\), we take \(\beta_0^* = 1\), which yields the highest joining rate. This can also be viewed as the limiting case of the equilibrium corresponding to \(v_0 (R, \mu) = \epsilon > 0\) when \(\epsilon\) is arbitrarily small. For any pair \((R, \mu)\) for which \(v_0 (R, \mu) < 0\), no queue will ever build up, irrespective of the strategy of the customers arriving at a busy expert. As a result, the expert would never make profits in the long run. We therefore restrict our analysis to all \((R, \mu)\) for which \(v_0 (R, \mu) \geq 0\).

The following Lemma characterizes the equilibrium joining decision when the expert is busy:

Lemma 3 Let

\[
B (R, \mu) = \frac{\mu_1}{\Lambda} \left( 1 - \frac{c}{\mu_1} \left( \frac{1}{v_1 (R, \mu)} - \frac{1}{\mu} \right) \right). 
\]

(4)

For a given \((R, \mu)\) and \(v_0 (R, \mu) \geq 0\), \(\beta^* (R, \mu) = (1, \beta_1^* (R, \mu))\) is uniquely determined as follows:

\[
\beta_1^* (R, \mu) = \begin{cases} 
1 & \text{if } B (R, \mu) > 1 \\
B (R, \mu) & \text{if } B (R, \mu) \in [0, 1] \\
0 & \text{if } B (R, \mu) < 0.
\end{cases}
\]

We introduce \(W_1 (\beta_1, \mu)\) as a shorthand notation for \(W ((1, \beta_1), \mu)\). Figure 1 illustrates the equilibrium \(\beta_1^*\): Setting the strategy of all other customers to \((1, \beta_1)\), the expected utility of a randomly arriving customer arriving when the expert is busy is plotted on the left panel. As the joining probability of all other customers increases, the expected waiting cost of the randomly arriving customer increases and his utility from joining monotonically decreases (as can also be verified from Equation 3). The best reaction of this customer is plotted on the right panel. It is to always enter for \(\beta_1 < B (R, \mu)\), where \(V_1 ((1, \beta_1), R, \mu) > 0\), or
Figure 1: Illustration of the equilibrium $\beta^*_1 = B(R, \mu)$ for $V = 1$, $\mu_1 = 1$, $\mu_0 = 0.95$, $A = 1$, $R = 0.15$, $r = 0.5$ and $c = 0.05$.

equivalently, where $v_1(R, \mu) > cW_1(\beta_1, \mu)$, i.e. where the expected net value of service is larger than the expected queuing cost. The best reaction is to randomize on $[0, 1]$ at $\beta_1 = B(R, \mu)$, where $v_1(R, \mu) = cW_1(B(R, \mu), \mu)$, and not to enter on $\beta_1 > B(R, \mu)$, where $v_1(R, \mu) < cW_1(\beta_1, \mu)$.

By definition, $(1, \beta^*_1)$ is an equilibrium if $(1, \beta^*_1)$ is a best response of a randomly arriving customer when all other customers use strategy $(1, \beta^*_1)$. As $v_0(R, \mu) \geq 0$, we already obtain with Lemma 2 that the best reaction of a customer arriving at an idle expert is to join; $\beta^*_0 = 1$. On Figure 1, $\beta^*_1(R, \mu)$ is identified by the intersection of the 45 degree line with the best response function $BR_1(\beta_1; R, \mu)$. Consequently, when $0 < B(R, \mu) < 1$, $\beta^*_1 = B(R, \mu)$, the randomization probability when arriving at a busy expert that makes the expected queuing cost equal to the net utility of service. If $B(R, \mu) < 0$, $V_1((1, \beta_1), R, \mu) < 0$ for all $\beta_1 \in [0, 1]$ and not joining is the best response to $\beta_1 = 0$ (no other customer joining), so, $\beta^*_1 = 0$ in this case. If $B(R, \mu) > 1$, $V_1((1, \beta_1), R, \mu) > 0$ for all $\beta_1 \in [0, 1]$ and joining is the best response to $\beta_1 = 1$ (all other customers joining), so, $\beta^*_1 = 1$ in this case. This completes the interpretation of Lemma 3.

From these results and the choice that $\beta^*_0 = 1$ when $v_0(R, \mu) = 0$, there are three relevant forms of $\beta^*$: $(1, 0)$, $(1, B(R, \mu))$ and $(1, 1)$. Figure 2 illustrates the regions in the $(R, r)$ space that give rise to these forms of $\beta^*$ for a fixed service strategy $\mu$. In the left panel of the top row, $\mu_0 = \mu_1$, whereas in all other panels, $\mu_0 < \mu_1$. On all panels, three lines are drawn: $L0 (v_0(R, \mu) = 0)$, $L1 (v_1(R, \mu) = cW_1(0, \mu))$ and
In the region \( \Omega_0 (\mu) \), the equilibrium joining strategy is \( \beta^* = (1, 0) \), i.e., customers do join an idle expert, but not a busy one. The service prices are so high that even if no other customer would join the busy expert, a randomly arriving customer would not join the busy expert either. In this region, the choice of \( R \) “prices out the busy expert” and nobody joins the busy expert. In the region \( \Omega_1 (\mu) \), the equilibrium joining strategy is \( \beta^* = (1, B (R, \mu)) \). Customers arriving at a busy expert join with probability \( B (R, \mu) \), but receive 0 utility from doing so. We call this “intermediate pricing.” In the region \( \Omega_2 (\mu) \), the equilibrium joining strategy is \( \beta^* = (1, 1) \). The service prices are so “cheap” that a randomly arriving customer obtains positive surplus from joining the expert in either state even when all other arriving customers join. Any vector outside \( \Omega_0 (\mu) \cup \Omega_1 (\mu) \cup \Omega_2 (\mu) \) has \( \beta_0^* = 0 \), i.e., no customer ever joins the expert.

Assumption 1 guarantees that the regions \( \Omega_0 (\mu) \) and \( \Omega_1 (\mu) \) are non-empty as their common border \( L1 \) exists in \( \mathbb{R}^2_+ \): \( v_1 (0, \mu) - cW_1 (0, \mu) = V - c \left( \frac{1}{\mu_1} + \frac{1}{\mu_0} \right) \geq V - \frac{2c}{L} > 0 \). However, Assumption 1 does not guarantee that \( \Omega_2 (\mu) \) is non-empty. When \( v_1 (0, \mu) - cW_1 (1, \mu) = V - c \left( \frac{1}{\mu_1} + W_1 (1, \mu) \right) < 0 \), line \( L2 \) lies in the negative quadrant of \( \mathbb{R}^2 \). In that case, \( \Omega_2 (\mu) \) is empty.

In \( \Omega_1 (\mu) \), all price vectors \( R \) for which the expected payment when joining a busy expert, \( R + \frac{c}{\mu_1} \), is constant result in the same equilibrium strategy \( B (R, \mu) \). This can be seen by observing that in \( B (R, \mu) \), the prices appear only via the term \( R + \frac{c}{\mu_1} \) (see Equation 4). The fixed fee and billing rates are thus substitutes in \( \Omega_1 (\mu) \): In order to obtain the same joining probability when the billing rate increases, the fixed fee needs to decrease. Of course, the billing rate is upper bounded by \( v_0 (R, \mu) \geq 0 \) (Equation 1) or \( R + \frac{c}{\mu_0} \leq V - \frac{2c}{L} \), since otherwise no customer would join the idle expert. When \( \mu_0 = \mu_1 \), this constraint is automatically satisfied as \( v_1 (R, \mu) - cW_1 (\beta_1, \mu) \leq v_0 (R, \mu) \) and \( v_1 (R, \mu) = v_0 (R, \mu) \); any customer who joins a busy expert would also join an idle expert.
Figure 2: Equilibrium regions for $V=1$, $\mu_1 = 1, \Lambda = 0.6, c = 0.125$ and $\mu_0 = 1, 0.90, 0.65$ and 0.55. In region $\Omega_0 (\mu)$, the equilibrium joining strategy is $\beta^* = (1,0)$. In region $\Omega_1 (\mu)$, the equilibrium joining strategy is $\beta^* = (1, B (R, \mu))$. In region $\Omega_2 (\mu)$, the equilibrium joining strategy is $\beta^* = (1,1)$. Outside $\Omega_0 (\mu)$, $\Omega_1 (\mu)$ and $\Omega_2 (\mu)$ customers do not join.

### 4.3 Optimal prices and profits for a given service strategy

In the previous section, we discussed the customer response for a given $(R, \mu)$. We now let the expert optimize over $R$ but keep $\mu$ fixed. Our approach is to find the locus of un-dominated $R$ pairs for $\Omega_0 (\mu)$, $\Omega_1 (\mu)$ and $\Omega_2 (\mu)$ and search in these sets for the optimal prices yielding each equilibrium type. We then compare the optimal profits of these three cases to find the optimal profit as a function of $\mu$.

**Dominating set of prices when pricing out the busy expert** $(R \in \Omega_0 (\mu))$: Recall that $\beta^* = (1, 0)$ in $\Omega_0 (\mu)$. We introduce $SW_1 (\beta_1, \mu)$ as a shorthand notation for $SW ((1, \beta_1), \mu)$. The expert’s optimal profits yielding equilibrium $(1,0)$ are characterized in the following Proposition:
Proposition 4 In $\Omega_0(\mu)$, the maximum profit rate of the expert, denoted by $\Pi_0(\mu)$, equals $SW_1(0, \mu)$. All price pairs satisfying

$$v_0(R, \mu) = 0 \text{ and } v_1(R, \mu) \leq cW_1(0, \mu)$$

achieve this profit rate.

In the region $\Omega_0(\mu)$, the expert prices out customers that arrive at a busy expert; customers only join when the expert is idle. As a result, all joining customers are homogeneous, i.e. they have the same utility net of prices, $v_0(R, \mu)$. This allows the expert to extract all surplus from the customers by setting prices that result in $v_0(R, \mu) = 0$ or $R + \frac{v_0}{\mu_0} = V - \frac{v_0}{\mu_0}$. Since the expert can use two prices (fixed fee and billing rate) to extract the customer surplus, multiple combinations of billing rates and fixed fees all extract the customer surplus; this is the $v_0(R, \mu) = 0$ ($L0$) boundary of the region. As the expert extracts all customer surplus, he earns the social surplus, $SW_1$, as profit.

**Dominating set of ‘intermediate’ prices ($R \in \Omega_1(\mu)$):** Recall that $\beta^* = (1, B(R, \mu))$ when $B(R, \mu) \in [0, 1]$. Define the interval $I(\mu) \doteq [0, \min(1, B(0, \mu))]$. For all $\beta_1 \in I(\mu)$, there exist price pairs in $\mathbb{R}^2_+$ that yield $(1, \beta_1)$ as an equilibrium. Recall from Section 4.2 that with Assumption 1, $\Omega_1(\mu)$ is always non-empty. Therefore, $I(\mu) \neq \emptyset$. The expert’s optimal profits over the set of prices yielding a given equilibrium $\beta_1 \in I(\mu)$ in conjunction with service strategy $\mu$ are characterized in the following Proposition:

**Proposition 5** Define $\Delta_0(\beta_1, \mu) = c\mu W_1(\beta_1, \mu) - (\mu_1 - \mu_0)V$ and let $\overline{B}(\mu)$ solve $\Delta_0(\beta_1, \mu) = 0$ for $\beta_1$.

In $\Omega_1(\mu)$, the maximum profit rate of the expert over all price pairs yielding equilibrium $\beta_1$, denoted by $\Pi_1(\beta_1, \mu)$ is as follows:

$$\Pi_1(\beta_1, \mu) = \begin{cases} 
SW_1(\beta_1, \mu) & \text{if } \beta_1 \in [0, \overline{B}(\mu)] \cap I(\mu) \\
SW_1(\beta_1, \mu) - L_1(\beta_1, \mu) & \text{if } \beta_1 \in [\overline{B}(\mu), 1] \cap I(\mu)
\end{cases}$$

where

$$L_1(\beta_1, \mu) = \Delta_0(\beta_1, \mu) \Lambda_0((1, \beta_1), \mu).$$
The price pairs that achieve these maximum profits are:

\[
\begin{align*}
R_I (\beta_1, \mu) &= \begin{cases} 
R_I (\beta_1, \mu) = -\frac{\mu_0}{\mu_1} \Delta_0 (\beta_1, \mu), & \text{if } \beta_1 \in [0, \overline{\beta} (\mu)] \cap I (\mu) \text{ "case I"} \\
\frac{w_1 (\beta_1, \mu)}{\mu_1} - 1, & \text{if } \beta_1 \in [\overline{\beta} (\mu), 1] \cap I (\mu) \text{ "case II"}
\end{cases}
\end{align*}
\]

\[
\begin{align*}
R_{II} (\beta_1, \mu) &= \begin{cases} 
R_{II} (\beta_1, \mu) = 0, & \text{if } \beta_1 \in [\overline{\beta} (\mu), 1] \cap I (\mu) \text{ "case II"}
\end{cases}
\end{align*}
\]

and constitute the locus of un-dominated price pairs for \( \beta_1 \in I (\mu) \).

If \( \mu_1 = \mu_0 \), then, \( \overline{\beta} (\mu) = -\infty \) (case I does not apply) and \( R_{II} (\beta_1, \mu) = \{ \mu \in \mathbb{R}^2_+ : v_1 (R, \mu) = cW_1 (\beta_1, \mu) \} \) is the locus of un-dominated price pairs for \( \beta_1 \in I (\mu) \).

Figure 3 illustrates Proposition 5. In the Figure, the locus of the un-dominated price pairs as \( \beta_1 \) ranges from 0 to 1 are highlighted with a thick line. The parameters used for Figure 3 are the same as for Figure 2.

First, we discuss the case when \( \mu_0 = \mu_1 \), where the expert does not induce service. Then, \( \overline{\beta} (\mu) = -\infty \) and only case II applies and \( \Pi_1 (\beta_1, \mu) = SW_1 (\beta_1, \mu) - L_1 (\beta_1, \mu) \). We refer to the difference between the social welfare and the expert’s profit rate at the ‘loss rate,’ which for \( \mu_0 = \mu_1 \) can be written as \( L_1 (\beta_1, \mu) = cW_1 (\beta_1, \mu) \Delta_0 ((1, \beta_1), \mu) \). Although a customer arriving at an idle expert does not incur any waiting costs, the expert is unable to extract all surplus from that customer when he does not induce service.

The reason is that the expected payment of the customer is independent of the state of the expert upon arrival and therefore, the expert needs to set the payment such that the customers arriving at a busy expert will join. This limits him to setting prices s.t. \( R + \frac{\mu}{\mu_1} = V - c \left( \frac{1}{\mu_1} + W_1 (\beta_1, \mu) \right) \); i.e., the expert cannot extract \( c \left( \frac{1}{\mu_1} + W_1 (\beta_1, \mu) \right) \) from customer value \( V \), irrespective of the state of the expert upon arrival. Defining \( \Lambda_1, \phi (\beta_1, \mu) = \Delta_0 ((1, \beta_1), \mu) + \Lambda_1 ((1, \beta_1), \mu) \), which is the total net arrival rate, the expert’s profit rate \( SW_1 (\beta_1, \mu) - L_1 (\beta_1, \mu) \) can be re-written as:

\[
\Pi_1 (\beta_1, \mu) = \left( V - c \left( \frac{1}{\mu_1} + W_1 (\beta_1, \mu) \right) \right) \Lambda_1, \phi (\beta_1, \mu).
\]

(6)

Next, we discuss case I when \( \mu_0 < \mu_1 \), where the expert’s profit rate is equal to the social welfare. Any price vector \( R \) in \( \Omega_1 (\mu) \) already extracts the surplus of a customer arriving when the expert is busy since \( v_1 = cW_1 \) in \( \Omega_1 (\mu) \). If the expert can select a price vector, \( R_I (\beta_1, \mu) \) such that \( v_0 \) is also 0, then, the expert extracts the surplus of the customer arriving when the expert is idle, too. In this case, the expert’s profit
rate is equal to the social welfare. This is the maximum profit rate that the expert can ever generate with service strategy $\mu$.

It is interesting that even with a differential service policy ($\mu_0 < \mu_1$), a fixed fee is not sufficient to extract all surplus as customers are not homogeneous: From Equation 3, it follows that the expected system waiting time of a customer arriving at a busy expert, $\frac{1}{\mu_1} + W_1(\beta_1, \mu)$, is always higher than the expected system waiting time of a customer arriving at an idle expert, $\frac{1}{\mu_0}$. With a fixed fee, the expert can extract the surplus from the latter, but not from the former. As a result, the customer arriving when the expert is idle always has positive surplus. With a differential service policy ($\mu_0 < \mu_1$) and a positive billing rate, the expert can extract more billing revenues from the customers that arrive when he is idle. This erodes the gain in surplus of the customers that arrive when the expert is idle. A positive fixed fee, a positive billing rate and a differential service policy may even allow the expert to capture all customer surplus and pocket the whole social welfare as profits, despite the heterogeneity of customers. That is why we refer to selecting a slower service rate for customers that arrive when the expert is idle as ‘skimming’. The locus of price pairs $R_I(\beta_1, \mu)$ that allow the expert to extract all customer surplus as $\beta_1$ varies is indicated as the thick line on $L0$ in Figure 3, bottom left and right panels.

Now, we discuss case II when $\mu_0 < \mu_1$, where the expert is unable to charge a price that extracts all surplus (i.e. that is on $v_0 = 0$) and leads to equilibrium $(1, \beta_1)$; this happens for large $\beta_1$. As the joining probability, $\beta_1$, increases, the waiting cost advantage of a customer arriving at an idle expert, becomes larger. The billing rate needed to extract the surplus of this customer thus increases. Recall from Section 4.2 that the billing rate and the fixed fee are substitutes. A large enough billing rate to extract the customer surplus for a given $\beta_1$ may require a negative fixed fee. Thus, there exists a threshold $\overline{B}(\mu)$ on $\beta_1$ above which the expert cannot extract all surplus with non-negative fixed fees. As we disallow negative fixed fees, the best the expert can do for a given $\beta_1$ that is higher than $\overline{B}(\mu)$ is to set the fixed fee equal to zero$^1$ and charge a billing rate that extracts all surplus of customers arriving when he is busy ($v_1 = 0$). The locus of price vectors $R_{II}(\beta_1, \mu)$ that achieve this for $\beta_1 \in [\overline{B}(\mu), 1]$ is indicated as the thick line on the vertical axis of

$^1$If a fixed subsidy of at most $S > 0$ would be allowed to entice customers to join, a similar structure would hold with $R_{II} = -S$. As we set $R \in \mathbb{R}_2^+$, we exclude this possibility by design. Expanding the price domain to finite negative prices (i.e. allowing for subsidies) would not change the insights of the paper.
Figure 3: Un-dominated price pairs in \( \Omega_1 (\mu) \), with the same parameters as in Figure 2. On the top left panel, the profits are identical on all lines parallel to \( R + \frac{z}{\beta} \); only five lines corresponding to five different values of \( \beta_1 \) are plotted. On the top right panel, \( B (\mu) < 0 \) and the un-dominated price pairs are all on the vertical axis, where \( R = 0 \). On the bottom left panel, \( B (\mu) \in (0, 1) \) and the un-dominated prices are on the vertical axis and on the line \( v_0 (R, \mu) = 0 \). On the bottom right panel, \( B (\mu) > 1 \) and the un-dominated price pairs are on \( v_0 (R, \mu) = 0 \). In the last three panels, \( \beta_1 \) ranges from 0 to 1 on the thick lines tracing out the un-dominated price vectors.

Figure 3, top right and bottom left panels.

**Dominating set of ‘cheap’ prices (R \( \in \Omega_2 (\mu) \)):** Recall that \( \beta^* = (1, 1) \) in \( \Omega_2 (\mu) \). The expert’s optimal profits are characterized in the following Proposition:

**Proposition 6** In \( \Omega_2 (\mu) \), if it is non-empty, the maximum profit rate of the expert is \( \Pi_1 (1, \mu) \).
Recall from section 4.2 that when \( V - c \left( \frac{1}{\mu_1} + W_1(\beta_1, \mu) \right) < 0 \), then \( \Omega_2 = \emptyset \) and the optimization over \( \Omega_2 \) becomes irrelevant. However, even when \( \Omega_2 \neq \emptyset \), the optimal price vector cannot lie inside \( \Omega_2 \), but, lies on the border with \( \Omega_1(\mu) \). The intuition is the following: As long as prices are in the interior of \( \Omega_2 \), the customer surplus is strictly positive. The expert can keep all customers joining while increasing the prices. On the \( L_2 \) boundary of the region, if it exists (as in Figure 7, bottom right panel), the profits decrease in \( R \) (or increase in \( r \)). On the \( L_1 \) boundary of the region, the profits increase in \( R \). Therefore, the profit-maximizing point in \( \Omega_2 \) is always the lowest-\( R \) point on \( L_2 \) that is contained in \( \Omega_2 \). This point is also in \( \Omega_1(\mu) \) and has \( \beta_1 = 1 \), so it is associated with an expert profit rate of \( \Pi_1 \), captured in Proposition 5.

As a result, region \( \Omega_2 \) does not need to be considered by the expert when determining the optimal pricing strategy for a given service rate, \( \mu \).

**Summary of optimal price conditions for a fixed service policy:** Now, we can search over all prices in \( \Omega_0(\mu) \cup \Omega_1(\mu) \cup \Omega_2(\mu) \). Combining Propositions 4, 5 and 6, we obtain that the expert’s optimal profit rate as a function of the service rate, denoted by \( \hat{\Pi}(\mu) \), is given by

\[
\hat{\Pi}(\mu) = \max_{\beta_1 \in I(\mu)} \begin{cases} 
SW_1(\beta_1, \mu) & \beta_1 \in \{ [0, \overline{B}(\mu)] \cap I(\mu) \} \cup \{0\} 
SW_1(\beta_1, \mu) - L_1(\beta_1, \mu) & \beta_1 \in [\overline{B}(\mu), 1] \cap I(\mu)
\end{cases}
\]  

(7)

The optimal profit \( SW_1(0, \mu) \) obtained in region \( \Omega_0(\mu) \) is subsumed in \( SW_1(\beta_1, \mu) \) when \( \beta_1 \) is allowed to take the value 0. To make sure \( \beta_1 = 0 \) is included in the search set even when \( \overline{B}(\mu) < 0 \), we use the union with the singleton \( \{0\} \) in case I of Equation 7.

**Proposition 7** The maximizer of (7), \( \beta_1^*(\mu) \), is either one of the end points of \( I(\mu) \), or one of the following interior maxima:

(I) the unconstrained maximizer of \( SW_1(\beta_1, \mu) \), denoted by \( \beta_1^I(\mu) \),

(II) the unconstrained maximizer of \( SW_1(\beta_1, \mu) - L_1(\beta_1, \mu) \), denoted by \( \beta_1^{II}(\mu) \),

(III) \( \overline{B}(\mu) \).

\( \beta_1^I(\mu) \) and \( \beta_1^{II}(\mu) \) are the root of the first order conditions of \( SW_1(\beta_1, \mu) \) and \( SW_1(\beta_1, \mu) - L_1(\beta_1, \mu) \) respectively that satisfy \( \beta_1^I(\mu) \Lambda < \mu_1 \) and \( \beta_1^{II}(\mu) \Lambda < \mu_1 \) respectively.

Proposition 7 establishes that the optimization problem of Equation 7 in \( \beta_1 \) is unimodal and therefore, either an interior maximizer, or a corner point are optimal. Based on Proposition 7, Figure 4 illustrates
how the \((c, \Lambda)\) space is partitioned into different regions that determine the expert’s optimal price structure, corresponding joining strategy and profit rate for a given \(\mu\). Observe that for large waiting costs, irrespective of the service strategy, the expert prices out the busy expert (region (c)). For small waiting costs, the price structure depends on the service strategy and the arrival rate (market potential). For \(\mu_0 = \mu_1\) (left panel), the social welfare cannot be extracted by the expert. In regions (a) and (b), the prices are \(R_{II}(\beta_1, \mu)\) and customer strategies are \((1, 1)\) and \((1, \beta_{II}^1)\) respectively. For \(\mu_0 < \mu_1\) (middle and right panel), there are parameters for which the social welfare can be extracted by the expert with prices \(R_I(\beta_1, \mu)\) (regions d, e and f). The customer strategies in these regions are \((1, \bar{B}(\mu))\), \((1, \beta_1^I)\) and \((1, 1)\) respectively.

5 Does it pay to induce service?

Why slow service rates may increase profits. So far, we have analyzed the optimal price and the corresponding customer equilibrium for a given service strategy \(\mu = (\mu_0, \mu_1)\). Now, we want to study which service policy \(\mu \in \{\mu_0, \mu_1\} \times \{\mu_0, \mu_1\}\) will maximize the expert’s profit rate. Recall from Equation 7 that the expert’s optimal profit rate yielding \((1, \beta_1)\) and \(\mu\) is either in the form of \(SW_1(\beta_1, \mu)\) or \(SW_1(\beta_1, \mu) - L_1(\beta_1, \mu)\). We begin by determining how this profit rate depends on the service rates:

**Proposition 8** The expert’s optimal profit rate yielding \((1, \beta_1)\) and \(\mu\) depends as follows on the service rates:

(i) \(\frac{\partial SW_1(0, \mu)}{\partial \mu_0} > 0\), \(\frac{\partial SW_1(0, \mu)}{\partial \mu_1} = 0\),

(ii) \(\frac{\partial SW_1(\beta_1, \mu)}{\partial \mu_0} > 0\), \(\frac{\partial SW_1(\beta_1, \mu)}{\partial \mu_1} > 0\) for any \(0 < \beta_1 \Lambda < \mu_1\),

(iii) \(\frac{\partial (SW_1(\beta_1, \mu) - L_1(\beta_1, \mu))}{\partial \mu_0} \leq 0\), \(\frac{\partial (SW_1(\beta_1, \mu) - L_1(\beta_1, \mu))}{\partial \mu_1} > 0\) for any \(0 < \beta_1 \Lambda < \mu_1\).

Parts (i) and (ii) of this Proposition are intuitive: The social welfare generated by the system is non-decreasing in the service rate. However, when the expert optimally does not extract all customer surplus, the expert’s profit rate yielding \(\beta_1\) for a given \(\mu\) may decrease in \(\mu_0\). This is part (iii). With Equation 7, the above observations lead to the following two Corollaries about \(\frac{\partial \Pi_i(\mu)}{\partial \mu_i}\) for \(i = 0\) and 1:

**Corollary 9** \(\frac{\partial \Pi_i(\mu)}{\partial \mu_0} \leq 0\): It may be optimal for the expert to select either a high service rate or a low service rate for customers arriving when he is idle.
Figure 4: Characterization of different regions that determine the form of $\hat{\Pi}(\mu)$ for $\mu_1 = 1$, with $\mu_0 = 1$ (left), $\mu_0 = 0.825$ (middle) and $\mu_0 = 0.725$ (right). With these parameters, the regions for the honest policy ($\mu_0 = \mu_1$) are depicted on the left panel. The regions for two different skimming policies ($\mu_0 < \mu_1$) are depicted on the middle and right panels. The table below indicates the pricing strategies, customer equilibrium and expert’s profit rate for each of the regions.

The intuition for this Corollary is that as the service rate $\mu_0$ decreases, the expert loss rate, $L_1$, may decrease faster than the loss in social welfare, leaving the expert better off. In this case, the expert may want to provide slow service to customers arriving when he is idle. The Corollary captures the key trade-off of service inducement and is the main driver of the results in our paper: Through service inducement via skimming, on the one hand, the expert reduces the ‘total pie’, but, on the other hand, he can capture a larger piece of it. Service induction will occur when capturing a larger piece of a smaller pie (with service inducement) is more profitable than capturing a smaller piece of a larger pie (with honest service).
Corollary 10 $\frac{\partial \Pi (\mu)}{\partial \mu_1} \geq 0$: It is optimal for the expert to select the high service rate for customers arriving when he is busy.

The Corollary is due to the fact that the monopoly profits always increase in $\mu_1$ in parallel with the increase in social welfare. Therefore, $\mu_1^* = \overline{\mu}$ and the two policies whose profits we need to compare to identify the optimal $\mu$ are $\mu_s = (\overline{\mu}, \overline{\mu})$ and $\mu_h = (\overline{\mu}, \overline{\mu})$, the service strategy for which the expert always induces service $(\overline{\mu}, \overline{\mu})$ and the the service strategy for which the expert induces service only when he is busy; $(\overline{\mu}, \overline{\mu})$, can be eliminated. It is the purpose of this section to determine when the skimming strategy $\mu_s$ will be more profitable for the expert than the honest strategy $\mu_h$. In the remainder of this section, we obtain analytical results for low waiting cost and present numerical results for general waiting cost.

Analytical insights: We use $c' = \frac{c}{\Lambda}$, the normalized waiting cost, $\rho = \frac{\Lambda}{\overline{\mu}}$, the base load and $\alpha = \frac{\mu}{\overline{\mu}}$, the service inducement potential, where a lower $\alpha$ means a higher induction potential. The following Proposition identifies limiting structural properties of the maximizer of Equation 7, $\beta_1^* (\mu)$, based on Proposition 7:

**Proposition 11** When $c' > 0$ but small,

if $\sqrt{c'} < 1 - \rho + O \left( (1 - \rho)^{\frac{3}{2}} \right)$, then $\beta_1^* (\mu_h) = 1$ and $\beta_1^* (\mu_s) = 1$; (8)

if $\sqrt{c'} > 1 - \rho + O \left( (1 - \rho)^{\frac{3}{2}} \right)$, then $\beta_1^* (\mu_h) = \beta_1^* (\mu_h) \in (0, 1)$ and $\beta_1^* (\mu_s) = \beta_1^* (\mu_s) \in (0, 1)$. (9)

Going back to Figure 4, Proposition 11 says that in the first case (Equation 8), regions (a) and (f), and in the second case (Equation 9), regions (b) and (e) apply under the honest and skimming strategies, respectively. Also, the boundary between regions (a) and (b) and between (e) and (f) can be approximated by $\sqrt{c'} = 1 - \rho$. Now, we can discuss when service inducement is profitable for the two cases in Proposition 11.

In the first case (Equation 8 of Proposition 11), the profit difference $\hat{\Pi} (\mu_h) - \hat{\Pi} (\mu_s)$ is $SW_1 (1, \mu_h) - L_1 (1, \mu_h) - SW_1 (1, \mu_s)$, which can be rewritten as

$$\hat{\Pi} (\mu_h) - \hat{\Pi} (\mu_s) = - \frac{\Lambda}{\text{arrival rate}} \left( \frac{1}{\Lambda} + W_1 (1, \mu_h) - \frac{Q_1 (1, \mu_h)}{\text{average waiting time}} \right) c,$$  (10)

where $Q_1 (\beta_1, \mu) = Q ((1, \beta_1), \mu)$. For $c = 0$, we obtain immediately from Equation 10 that $\hat{\Pi} (\mu_s) = \hat{\Pi} (\mu_h)$; when congestion cost does not matter, the expert is indifferent. For $c > 0$, the profitability of service
inducement is determined by the waiting time in the system conditional on the expert being busy upon arrival with the honest service strategy vs. the average waiting time in the system with the skimming service strategy. This can be understood as follows: Recall from Equation 6 that with honest service, the expert through its pricing strategy is unable to extract from the generated service value per customer, \( V \), the expected system waiting cost conditional on arriving at a busy expert, \( c \left( \frac{1}{\rho} + W_1(1, \mu_h) \right) \). With skimming, the expert appropriates the whole customer surplus. Therefore, the expert is only unable to extract from the generated service value per customer, \( V \), the average waiting cost per customer, \( c Q_{(1, \mu_s)} \). As all customers join with either service strategy, the value generated is \( V \Lambda \) per unit of time for both service strategies. Therefore, the profit difference reduces to the waiting cost difference, as expressed in Equation 10.

From the expressions derived in Proposition 1, we characterize the difference between the profits under the skimming and honest services strategies:

**Proposition 12** If \( c' \) is small and \( \sqrt{c'} < 1 - \rho + O \left( (1 - \rho) \frac{1}{2} \right) \), then \( \hat{\Pi}(\mu_h) = \hat{\Pi}(\mu_s) \) for \( c' = 0 \), otherwise

\[
\hat{\Pi}(\mu_h) - \hat{\Pi}(\mu_s) = -\frac{1 - 2\alpha - (2 - \alpha - \frac{1}{\alpha})\rho}{1 + \alpha \frac{\rho}{1 - \rho}} c' V \Lambda, \quad \text{and} \quad (11)
\]

(i) for \( \alpha < \frac{1}{2} \), \( \hat{\Pi}(\mu_h) > \hat{\Pi}(\mu_s) \) \( \forall \rho \).

(ii) for \( \frac{1}{2} < \alpha < \frac{\sqrt{5} - 1}{2} \approx 0.618 \), \( \hat{\Pi}(\mu_s) > \hat{\Pi}(\mu_h) \) for \( \rho < \frac{1 - 2\alpha}{2 - \alpha - \frac{1}{\alpha}} \).

Proposition 12 allows us to understand the impact of the service inducement potential, \( \alpha \), and base load, \( \rho \), on the profit difference between the skimming and honest service strategies. As \( 2 - \alpha - \frac{1}{\alpha} < 0 \) for \( \alpha \in (0, 1) \), it can be seen in the numerator of Equation 11 that when \( \alpha < \frac{1}{2} \), \( \hat{\Pi}(\mu_h) > \hat{\Pi}(\mu_s) \) for any positive \( \rho \). This can be understood by studying Equation 10 when the base load tends to zero: The time in the system of an honest expert conditional on arriving when he is busy tends to the expected remaining service time for the customer currently in service, \( \frac{1}{\mu} \), plus the service time of the arriving customer, \( \frac{1}{\mu} \). The average waiting time in the system of a skimming expert tends to the average service time, which is \( \frac{1}{\mu} \) as mostly customers arrive when the expert is idle. Thus, with Equation 10, we obtain that when \( \frac{2}{\rho} < \frac{1}{2} \), or \( \alpha < \frac{1}{2} \), service inducement is never optimal for low values of \( c' \).

When \( \alpha > \frac{1}{2} \), the average waiting time when the expert skims is less than the waiting time when the honest expert is busy, i.e. service inducement may become profitable. It can be seen in Equation 11 that
\( \rho = \frac{1 - 2\alpha}{2 - \alpha} \) is the base load that makes \( \hat{\Pi}(\mu_h) = \hat{\Pi}(\mu_s) \) for \( c' > 0 \). Below the base load threshold, \( \rho \), service inducement is profitable. The threshold is independent of the waiting cost. Furthermore, \( \rho \) increases for \( \alpha > \frac{1}{2} \), i.e., as the service inducement potential decreases, the region where it is optimal to skim gets larger: Skimming works for more and more loaded systems as the slowdown due to \( \mu_s \) decreases. To analyze \( \alpha > \sqrt{5} - \frac{1}{2} \), we need to revert to the second case of Proposition 11.

In the second case (Equation 9 of Proposition 11), the expert’s profit difference \( \hat{\Pi}(\mu_h) - \hat{\Pi}(\mu_s) \) is \( SW_1(\beta_I(\mu_h), \mu_h) - L(\beta_I(\mu_h), \mu_h) - SW_1(\beta_I(\mu_s), \mu_s) \). \( \beta_I(\mu_h) \) and \( \beta_I(\mu_s) \) are obtained from Proposition 7. Even though we can obtain closed form expressions for all terms, the analysis of the profit difference paralleling Equation 10 is not straightforward. However, we obtain the following insights for small values of the waiting cost:

**Proposition 13** If \( c' \) is small and \( \sqrt{c'} > 1 - \rho + O\left(\left(1 - \rho\right)^{\frac{3}{2}}\right) \), then \( \hat{\Pi}(\mu_h) = \hat{\Pi}(\mu_s) \) for \( c' = 0 \), otherwise for \( \frac{\sqrt{5} - 1}{2} < \alpha < 1 \)

\[
\hat{\Pi}(\mu_h) - \hat{\Pi}(\mu_s) = \left(2\left(1 - \sqrt{\alpha + (1 - \alpha)\rho}\right)\sqrt{\frac{c'}{\rho}} + \left(2 - \frac{1}{\alpha}\right)\rho - (1 - \alpha)(\rho + 1)\right)\sqrt{\rho} + O\left(c'^{\frac{3}{2}}\right), \tag{12}
\]

and \( \hat{\Pi}(\mu_s) > \hat{\Pi}(\mu_h) \) for \( \frac{\sqrt{5} - 2\alpha}{1 - \alpha} \sqrt{c'} < 1 - \rho + O\left((1 - \rho)^{2}\right) \).

From Equation 12, we notice that the profit difference is determined by a term proportional to \( \sqrt{c'} \) and a term proportional to \( c' \). For \( c' = 0 \), we obtain immediately that \( \hat{\Pi}(\mu_s) = \hat{\Pi}(\mu_h) \), as in Proposition 12. It can be seen in Equation 12 that \( \rho = 1 - \frac{\sqrt{5} - 2\alpha}{1 - \alpha} \sqrt{c'} \) is the base load that (approximately) makes \( \hat{\Pi}(\mu_h) = \hat{\Pi}(\mu_s) \) for \( c' > 0 \). If the base load is less than \( \rho \), service inducement is profitable. Notice that, contrary to the previous case (Equation 11 of Proposition 12), for positive waiting costs, the threshold base load depends on the waiting cost. Furthermore, notice that if \( \frac{\sqrt{5} - 2\alpha}{1 - \alpha} > 0 \), or, \( \alpha < \frac{1}{\sqrt{5}} \approx 0.707 \), \( \rho \) decreases in \( c' \) and is less than one and for \( \alpha > \frac{1}{\sqrt{5}} \), \( \rho \) increases in \( c' \) and is more than one. In other words, as the service inducement potential decreases, the region where it is optimal to skim gets larger: skimming works for more and more loaded systems, even for systems with a base load higher than one if the service inducement potential is low enough, as the slowdown due to \( \mu_s \) decreases. Table 2 summarizes the insights derived in Propositions 12 and 13.
Table 2: Overview of the profitability of service induction as a function of the normalized waiting cost, $c'$, the base load, $\rho$, and the service induction potential, $\alpha$, based on Propositions 12 and 13. Service induction (through skimming) is profitable when $\hat{\Pi}(\mu_h) \leq \hat{\Pi}(\mu_s)$.

<table>
<thead>
<tr>
<th>waiting cost</th>
<th>base load</th>
<th>induction potential</th>
<th>profit comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c' = 0$</td>
<td>$\rho \geq 0$</td>
<td>$\alpha \in (0, 1)$</td>
<td>$\hat{\Pi}(\mu_h) = \hat{\Pi}(\mu_s)$</td>
</tr>
<tr>
<td>$c' &gt; 0$ and small</td>
<td>$\rho &lt; 1 - \sqrt{c'}$</td>
<td>$\alpha \in (0, \frac{1}{2})$</td>
<td>$\hat{\Pi}(\mu_h) \geq \hat{\Pi}(\mu_s)$</td>
</tr>
<tr>
<td></td>
<td>$\rho &lt; 1 - \sqrt{c'}$</td>
<td>$\alpha \in \left(\frac{1}{2}, \frac{\sqrt{5} - 1}{2}\right)$</td>
<td>$\hat{\Pi}(\mu_h) \leq \hat{\Pi}(\mu_s)$ if $\rho \leq \frac{1 - 2\alpha}{\frac{1}{2} - \alpha - \sqrt{c'}}$</td>
</tr>
<tr>
<td></td>
<td>$\rho &gt; 1 - \sqrt{c'}$</td>
<td>$\alpha \in \left(\frac{\sqrt{5} - 1}{2}, 1\right)$</td>
<td>$\hat{\Pi}(\mu_h) \leq \hat{\Pi}(\mu_s)$ if $\rho \leq 1 - \frac{\frac{1}{2} - 2\alpha}{\frac{1}{2} - \alpha - \sqrt{c'}}$</td>
</tr>
</tbody>
</table>

Numerical observations: A full analysis of when service inducement is optimal for any waiting cost requires comparing the relevant profit rate expressions for honest and skimming service for any given parameter set. As can be observed from Figure 4, comparing the left with the right and medium panels, there are different combinations possible for which one particular pricing strategy is optimal with honest service and another with skimming service. Fortunately, the analysis for low values of the waiting cost in Propositions 12 and 13 (i.e. comparing regions (a) with (f) and regions (b) with (e)) is quite insightful and captures most of the structural results. We complete the section with three representative numerical examples that we explain using the analytical insights developed in the previous discussion. Figure 5 plots the indifference curve $\hat{\rho}(c')$ in the $(c', \rho)$ space that satisfies $\hat{\Pi}(\mu_s) = \hat{\Pi}(\mu_h)$ for high ($\alpha = 0.55$), medium ($\alpha = 0.65$) and low ($\alpha = 0.75$) service inducement potential, respectively. The figure is obtained by comparing the relevant profit rate expressions of $\hat{\Pi}(\mu_s)$ and $\hat{\Pi}(\mu_h)$ obtained in section 4.2 for every $(c, \Lambda)$ tabulated in Figure 4. Notice that on the left panel, $\frac{1}{2} < \alpha < 0.618$. As predicted by Proposition 12, the upper threshold is in $(0, 1)$ and independent of $c'$. On the middle and the right panels, $0.618 < \alpha < 0.707$ and $0.707 < \alpha < 1$, respectively, such that the upper threshold in decreasing and increasing in $c'$ respectively, as predicted by Proposition 13. Figure 4 also indicates other thresholds on the region where service inducement is profitable. We discuss these thresholds in the following two observations. As can be seen on Figure 5:

**Observation 1** When the waiting cost is large, service inducement is never optimal.
Figure 5: For a given service policy, characterization of different regions where skimming is optimal in the \((\rho - c')\) space for \(\alpha = 0.55\) (left panel), \(\alpha = 0.65\) (middle panel) and \(\alpha = 0.75\) (right panel).

This observation can be understood as follows: From Figure 4, observe that pricing out the busy expert is optimal for large waiting costs with or without skimming (case (c) in Figure 7). Only customers finding the system idle enter. All joining customers are therefore homogeneous in their net utility of service and the expert can earn the social welfare, \(SW_1(0, \mu_h)\) and \(SW_1(0, \mu_s)\), by extracting all the surplus of entering customers setting prices that satisfy \(R + \frac{c'}{\mu_0} = V - \frac{c}{\mu_0}\), given his service rate \(\mu_0\). Slowing down the service rate makes all customers wait more during service while no value is added and thus reduces the social welfare and profits that the expert can earn, as is also stated in Proposition 8(i). This is why skimming is suboptimal when the waiting costs are large.

Observation 2 When the waiting cost is low or medium, service inducement is not profitable for very low arrival rates.

This observation can be understood as follows: From Figure 4, observe that pricing out the busy expert is optimal for medium waiting cost without skimming (case (c) in Figure 4, left panel) and that cheap pricing is optimal for medium waiting cost with skimming (case (f) in Figure 4, middle or right panel). The expert’s profit difference is \(\hat{\Pi}(\mu_h) - \hat{\Pi}(\mu_s) = SW_1(0, \mu_h) - SW_1(1, \mu_s)\). With both strategies, the expert can extract the social welfare. The underlying instruments are quite different: With honest service, the expert’s prices are so high that no customer joins when the expert is busy. Therefore, only customers join when he
is idle. The joining customer base is homogeneous and the expert can thus price such that he extracts all surplus of entering customers. With skimming, the expert extracts all surplus with low prices, attracting all customers. The customer base is heterogeneous, but the expert deploys its service strategy and two price instruments to differentially extract surplus from the customers arriving when he is busy vs. when he is idle. Thus, honest service results in low arrival rates of customers and a more efficient service strategy; skimming results in high net arrival rates of customers and a less efficient service strategy. As a result, when the base load is low, the advantage of skimming will not be sufficient to be more profitable than honest service.

**Summary:** In summary, we have obtained the following insights: The expert implements service inducement only through skimming. With skimming, the expert is able to appropriate a larger share of the generated social welfare than with honest service, but the trade-off is that the total generated social welfare is lower with skimming. Whether skimming or honest service dominates relates to the effect of congestion and the cost of congestion.

1. When the waiting cost is not high, service inducement is profitable when the service inducement potential is low enough and the base load is between upper and lower thresholds.

2. The upper threshold on the base load depends strongly on the service inducement potential, when the service inducement potential decreases, the region in which service inducement is profitable increases.

3. When the waiting cost is high, service inducement is never profitable.

### 6 Conclusion, Discussion and Further Research

In this paper, we analyze the optimal price structure and service inducement strategy of a monopolist expert who sells a credence service. In this setting, the expert may have an incentive to provide unnecessary service since it is expensive for the customer to contest this. In particular, with a variable-rate contract, the expert’s revenues increase as a function of the total service time, which makes inducement feasible. If in addition, the expert’s workload fluctuates, he may have a strong incentive to induce service when the workload is low.

We introduce a simple queuing model that captures the key workload dynamics. The model incorporates several important elements such as customer value of service, market potential, waiting cost and service time,
and is able to capture the important tradeoffs concerning service inducement in a dynamic environment. Within this framework, we determine the optimal policy for the expert as a function of the characteristics of the environment. We find that three parameters dictate the optimal policy: (1) the base load (normalized to the base service rate); (2) the waiting cost (normalized to the value contribution per unit of time with the base service rate) and (3) the opportunity for the expert to induce service, which is the ratio of the slow service rate to the base service rate. Our model allows us to study two levers that increase the revenues of a monopolist for credence services: The price structure and the service policy. Below, we describe the insights that we obtain from our analysis.

**Insights.** Our main contribution is to show that when workload dynamics due to stochastic arrivals and service times are taken into account, there are conditions under which the capacitated expert will find it optimal to induce service even when serving a homogeneous consumer base. This phenomenon derives from incorporating workload dynamics in the analysis: When the workload level upon arrival is observable to the customer, this is a source of effective customer heterogeneity despite the homogeneous nature of the customer base and makes inducement desirable for the expert. The strategic utilization of the service inducement policy and the pricing structure allows the expert to skim the surplus of customers arriving at low workload levels. We believe that this interpretation of service pricing is new, and is obtained by explicitly considering the operations aspect of service delivery.

With our model, we show that a necessary condition for service inducement to be profitable is that the waiting costs are not too high. We uncover the role of the opportunity for the expert to induce service. When the potential is very high, then service inducement is not profitable as the slowdown due to skimming creates too much waiting cost. When the potential decreases, then service inducement becomes profitable for base loads (ratio of the potential arrival rate over the base service rate) less than one. When the opportunity further decreases, service inducement becomes profitable for base loads of more than one.

**Design Levers to Limit Potential Service Inducement.** Service inducement can create negative publicity and result in high losses if detected. From a design perspective, it is therefore important for management to better understand when service inducement is likely to emerge within their organization. Certainly, a time-based fee (e.g. an hourly billing rate) in conjunction with unverifiable service rate is
conducive to service inducement. Nevertheless, our model shows that this structure alone does not result in the expert always inducing service. There is one case in particular where the expert has a strong incentive to induce service: this is when there is some, but not too much overcapacity and the waiting cost is low.

Our results provide some possibilities to limit inducement in this case. An obvious measure would be to restrict the price structure to one with a fixed fee only. However, this may be perceived as being unfair since customers who need only a low amount of service pay the same as customers who need a high level of service. Another measure would be to invest in expert training and an ethical work culture in environments where inducement is deemed to be more likely. Operational measures can also be considered. Customer heterogeneity, in this case induced by workload fluctuations, is what drives service inducement. Steps to decrease workload fluctuations by decreasing the variability of inter-arrival or service times would be effective in limiting the amount of service inducement. Similarly, if customers are a priori heterogeneous with respect to valuation or waiting cost, segmenting customers and assigning different servers to homogeneous consumer segments could be considered. Matching capacity with market demand can also be useful since buffer capacity may be used to induce service. Using effective workload planning models with the possibility of redeploying experts in other customer segments could be used to control the incentives to induce service.

**Discussion of Assumptions and Future Research Directions.**

We now discuss some findings from related research on service inducement. In Debo et al. (2004) and (2006) we analyzed different models in which service inducement may be optimal. We also indentified models in which it is easily shown that service inducement is not optimal. In particular, in Debo et al. (2004), we developed a game in which customers fully observe the queue length upon arrival and the expert induces service only on the customer that would leave the system empty. We found that service inducement through skimming may be optimal. If the workload is concealed, customers remain ex-post homogenous and service inducement never occurs; a fixed fee with no service inducement is optimal. In Debo et al. (2006), we developed a game in which the customers observe three states, idle, one customer and more than one customer and the expert determines the service rate after observing the customer’s joining strategy (instead of determining the service rate upon arrival as in our current model). We found that service inducement through skimming may be optimal and derived the conditions when this holds. It can easily be shown that
reducing the observability of the queue length in the model of Debo et al. (2006) to two states, idle or busy, would lead to skimming never being subgame perfect and make service inducement unprofitable. Thus, our research provides a comprehensive analysis of service inducement in queuing systems.

Nevertheless, we make several simplifying assumptions. Below, we discuss the implications of relaxing some of these assumptions. We assumed that the queue length is concealed. Since we showed that it is the ex-post heterogeneity of customers upon arrival (due to observing different system states) that makes it profitable for the expert to induce service, making the queue length fully observable would only enhance this result.

We assumed a homogeneous customer base and showed that even with a homogeneous customer base, due to ex-post heterogeneity, service inducement would be observed. If customer heterogeneity with respect to the valuation of the service and/or with respect to waiting costs were introduced into our model, we expect the level of service inducement to only increase.

We assumed that the cost of service per unit time is 0. With this assumption, service inducement only has an implicit cost to the expert in that it uses up limited capacity. If the service cost per unit time were positive, service inducement would become less attractive to the expert.

We assumed that the over-provision of service does not detract value. If the over-provision of some services can directly harm the consumer, such as with unnecessary medical intervention, this would destroy total surplus and reduce the ability of the expert to extract profit via service inducement.

We assumed that the expert is a monopolist. For specialized services, the entry cost to the market may be substantial, justifying the monopolist assumption in part. In the presence of multiple experts, customers may select the expert based on his/her workload upon arrival. Our conjecture is that the competitive pressure will reduce the profitability of service inducement. However, it is unclear whether competition will completely eliminate service inducement. For a model with endogenous entry of experts into the market, see Emons (1997). However, workload fluctuations are not discussed in his model.

In this paper, we developed a basic model in order to uncover how a phenomenon that is typically studied in the operations literature – workload fluctuation due to stochastic inter-arrival and processing times – influences an expert’s incentives to induce service. We hope that our work generates further research
in this area. Detailed and specific models of expert-customer interactions in different industries could be developed if the goal is to generate insights into a particular industry. In our model, the expert is the owner of the service company and chooses, as a monopolist, the optimal price structure. Kalra et al. (2003) study compensation schemes in a principal-agent model in which the agent has the opportunity to oversell services. It would be interesting to investigate design problems as in Kalra et al., but where the agent is a credence-good expert modeled as a single-server queue. As shown in the economics literature, competition and reputation are two important factors that limit service inducement. Analyzing these phenomena with the same expert server model may yield interesting insights into the interaction of workload fluctuation, competition and reputation. Finally, it would be interesting to investigate how the credence characteristic of a service impacts capacity investment decisions.

7 Acknowledgements

Laurens Debo wishes to thank the Sasakawa Young Leaders Fellowship Fund for financial support. The authors thank the reviewers and Associate Editor for their constructive and detailed comments that helped to improve the paper significantly.

8 References


9 Appendix

Throughout the Appendix, we assume that $\frac{\theta_1 \Lambda}{\mu_1} < 1$.

Proof of Proposition 1.
Overview: We prove the proposition by first developing a continuous time Markov Chain of a queue in which the service time distribution (exponential with rate \( \mu_i \)) is a function of the state of the queue upon arrival of the customer. There are two possible states: \( i = 0 \) indicates that the queue is empty upon arrival and \( i = 1 \) indicates that the queue is non-empty upon arrival. The long run probabilities that \( n \) customers are in the system are calculated in step I. This gives \( \pi_0 \). In step II, the expected queue length, \( Q \), and conditional waiting time in the queue given that the queue is non-empty upon arrival, \( W \), are calculated.

Step I: The long run probability that \( n \) customers are in the system. Let 0 be the state in which the expert is idle. \( n \) is the state in which \( n \geq 1 \) customers are in the queue, where the customer that is currently under service arrived when the expert was idle. \( n' \) is the state in which \( n \geq 1 \) customers are in the queue, where the customer that is currently under service arrived when the expert was not idle. Let \( \lambda_i = \Lambda \beta_i \) for \( i = 0, 1 \). A transition rate diagram is depicted in Figure 6. An arrival in state 0, call it customer0, takes the system to state 1. Subsequent arrivals during this customer’s service take the system to states 2, 3 etc.. When customer 0 leaves the system (with rate \( \mu_0 \)), the system moves from state \( n \) to state \( n - 1' \) because the customer now in service had arrived when the server was busy. Arrivals and departures then happen at rates \( \lambda_1 \) and \( \mu_1 \) until the system goes to state 0.

Step Ia: The continuous time Markov Chain equilibrium conditions. We first determine the long run equilibrium conditions. The transition probabilities are as follows:

\[
\lambda_0 \pi_0 = \mu_0 \pi_1 + \mu_1 \pi_1', \quad (13)
\]

\[
(\mu_0 + \lambda_1) \pi_1 = \lambda_0 \pi_0 \quad (14)
\]

and

\[
(\mu_1 + \lambda_1) \pi_1' = \mu_0 \pi_2 + \mu_1 \pi_2', \quad (15)
\]

Finally, for more than two customers in the queue:

\[
(\mu_0 + \lambda_1) \pi_n = \lambda_1 \pi_{n-1}, n \geq 2 \quad (16)
\]

\[
(\mu_1 + \lambda_1) \pi_{n'} = \mu_0 \pi_{n+1} + \mu_1 \pi_{n'+1} + \lambda_1 \pi_{n'-1}, n \geq 2 \quad (17)
\]

Step Ib: Solution of the continuous time Markov Chain. We now solve for \( \pi_n \) and \( \pi_{n'} \). From
Figure 6: Transition rate diagram for the Markov process of the service system, with $\lambda_0$ ($\lambda_1$) as the arrival rate when the expert is idle (busy) and $\mu_0$ ($\mu_1$) the service rate when the expert is idle (busy). The state $n$ ($n'$) indicates the number of customers in the system when the customer that is being served arrived when the expert was idle (busy). To understand this representation, note that the thick line indicates the changes of the state when a customer arrives at an idle expert, followed by two customers arriving during his service time. During the service time of the other two customers, no new customers arrive and the system empties again.

Equations (13) and (14), we obtain that

$$\pi_{1'} = \frac{\lambda_1}{\mu_0 + \lambda_1} \pi_0 \quad \text{and} \quad \pi_1 = \frac{\lambda_0}{\mu_0 + \lambda_1} \pi_0$$

We now conjecture that

$$\pi_n = A \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^{n-1} \pi_0, \ n \geq 2$$
and that

$$\pi_{n'} = \left( C \left( \frac{\lambda_1}{\mu_1} \right)^n + B \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^{n-1} \frac{\lambda_0}{\mu_0 + \lambda_1} \right) \pi_0, \ n \geq 2$$  \hspace{1cm} (20)$$

for some $A$, $B$ and $C$. We have thus unknowns $A$, $B$, $C$ and $\pi_0$.

**Step Ic-i:** determination of $A$. Substituting the first conjecture in Equation 16, we obtain $A = \frac{\lambda_0}{\mu_0 + \lambda_1}$.

**Step Ic-ii:** determination of $B$. Substituting the second conjecture in Equation 17, we obtain an expression for $B$ by comparing the terms with $\left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^n$:

$$B (\mu_1 + \lambda_1) \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^{n-1} \frac{\lambda_0}{\mu_0 + \lambda_1} = \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^n \left( \frac{\lambda_0}{\mu_0 + \lambda_1} \mu_0 + B \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \mu_0 + \lambda_1 \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^{-2} \frac{\lambda_0}{\mu_0 + \lambda_1} \right) \right)$$

$$\implies B = -\frac{\lambda_1}{\lambda_1 - \mu_1 + \mu_0}$$

**Step Ic-iii:** determination of $C$. It follows from Equation 20 for $n' = 2'$ that

$$\pi_{2'} = \left( C \left( \frac{\lambda_1}{\mu_1} \right)^2 + B - \frac{\lambda_1}{\mu_0 + \lambda_1} \frac{\lambda_0}{\mu_0 + \lambda_1} \right) \pi_0$$

and from Equation 17 that

$$(\mu_1 + \lambda_1) \pi_{2'} = \left( \mu_0 A \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^2 + \mu_1 \left( C \left( \frac{\lambda_1}{\mu_1} \right)^3 + B \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^2 \frac{\lambda_0}{\mu_0 + \lambda_1} \right) \right) \pi_0 + \lambda_1 \frac{\lambda_1 \lambda_0}{\mu_0 + \lambda_1} \pi_0$$

We can solve for $C$, as $A$ and $B$ have been obtained already. The result is:

$$\frac{\mu_0 A \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^2 + \mu_1 \left( C \left( \frac{\lambda_1}{\mu_1} \right)^3 + B \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^2 \frac{\lambda_0}{\mu_0 + \lambda_1} \right)}{\mu_1 + \lambda_1} \pi_0 = C \left( \frac{\lambda_1}{\mu_1} \right)^2 + B \frac{\lambda_1}{\mu_0 + \lambda_1} \frac{\lambda_0}{\mu_0 + \lambda_1}$$

$$\implies C = \frac{\lambda_0}{\lambda_1 + \mu_0 - \mu_1}$$

**Step Ic-iv:** determination of $\pi_0$. Finally, we obtain $\pi_0$ by summing up all probabilities

$$\sum_{n=2}^{\infty} (\pi_n + \pi_{n'}) = \sum_{n=2}^{\infty} \left( A \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^{n-1} + C \left( \frac{\lambda_1}{\mu_1} \right)^n + B \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^{n-1} \frac{\lambda_0}{\mu_0 + \lambda_1} \right) \pi_0, \ n \geq 2$$

$$= \left( \frac{\lambda_0}{\lambda_1 + \mu_0 - \mu_1} - \frac{\lambda_1}{\lambda_1 + \mu_0 - \mu_1} \sum_{n=0}^{\infty} \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^n + \frac{\lambda_0}{\lambda_1 + \mu_0 - \mu_1} \left( \frac{\lambda_1}{\mu_1} \right)^2 \sum_{n=0}^{\infty} \left( \frac{\lambda_1}{\mu_1} \right)^n \right) \pi_0$$

$$= \left( \frac{\lambda_0}{\lambda_1 + \mu_0 - \mu_1} - \frac{\lambda_1}{\lambda_1 + \mu_0 - \mu_1} \frac{1}{1 - \frac{\lambda_1}{\mu_0 + \lambda_1}} + \frac{\lambda_0}{\lambda_1 + \mu_0 - \mu_1} \left( \frac{\lambda_1}{\mu_1} \right)^2 \frac{1}{1 - \frac{\lambda_1}{\mu_1}} \right) \pi_0$$

$$= \frac{\lambda_0 \lambda_1}{\mu_0} \left( \frac{1}{1 - \frac{\lambda_1}{\mu_1}} + \frac{\mu_0}{\mu_1} + \frac{\lambda_1}{\mu_1} \right) \pi_0$$
First, we calculate:

**Step II: the expected number of customers in the system.**

From which follows that

\( n \) which, after some algebraic manipulations can be rewritten as

\[
\pi_0 = 1 + \frac{\lambda_1}{\mu_0} \frac{\lambda_0}{\mu_0 + \lambda_1} + \frac{\lambda_0 \lambda_1}{\mu_0 + \lambda_1} \left( \frac{\mu_0}{\mu_1} + \frac{\mu_0 \left( 1 - \frac{\mu_0}{\mu_1} \right)}{\mu_1} \right) \pi_0 = 1
\]

or

\[
\pi_0 = \frac{1 - \lambda_1}{1 + \frac{\lambda_0}{\mu_0} - \frac{\lambda_1}{\mu_1}}
\]

With these values of \( A, B, C \) and \( \pi_0 \), Equations 18-20 satisfy the conditions of Equations 13-17.

**Step II: the expected number of customers in the system.** Now, we can calculate

\[
Q = \sum_{n=1}^{\infty} n \left( \pi_n + \pi_n' \right)
\]

First, we calculate:

\[
\sum_{n=2}^{\infty} n \left( \pi_n + \pi_n' \right) = \left( \frac{\lambda_0}{\lambda_1 + \mu_0 - \mu_1} \sum_{n=2}^{\infty} n \left( \frac{\lambda_1}{\mu_1} \right)^n + \frac{\lambda_0}{\lambda_1 + \mu_0 - \mu_1} \sum_{n=2}^{\infty} n \left( \frac{\lambda_1}{\mu_0 + \lambda_1} \right)^{n-1} \right) \pi_0
\]

\[
= \left( \frac{\lambda_0}{\lambda_1 + \mu_0 - \mu_1} \left( \frac{2 - \frac{\lambda_1}{\mu_1}}{1 - \frac{\mu_0}{\lambda_1}} \right) + \frac{\lambda_0}{\lambda_1 + \mu_0} \left( \frac{1 + \frac{\lambda_1}{\mu_0} - \mu_1}{\lambda_1 + \mu_0 - \mu_1} \left( \frac{2 - \frac{\lambda_1}{\mu_0}}{\mu_0} \right) \right) \right) \pi_0
\]

and obtain

\[
Q = \sum_{n=1}^{\infty} n \left( \pi_n + \pi_n' \right) = \left( \frac{\lambda_1}{\mu_0 + \lambda_1} + \frac{\lambda_0}{\mu_0 + \lambda_1} + \sum_{n=2}^{\infty} n \left( \pi_n + \pi_n' \right) \right) \pi_0
\]

\[
= \left( \frac{\lambda_0}{\lambda_1 + \mu_0 - \mu_1} \left( \frac{2 - \frac{\lambda_1}{\mu_1}}{1 - \frac{\mu_0}{\lambda_1}} \right) + \frac{\lambda_0}{\lambda_1 + \mu_0} \left( \frac{1 + \frac{\lambda_1}{\mu_0} - \mu_1}{\lambda_1 + \mu_0 - \mu_1} \left( \frac{2 - \frac{\lambda_1}{\mu_0}}{\mu_0} \right) \right) \right) \pi_0
\]

which, after some algebraic manipulations can be rewritten as

\[
Q = \frac{\lambda_0}{\mu_0} \left( \frac{1}{\mu_1} + \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) + \lambda_1 \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) \lambda_1 = \frac{\lambda_0}{\mu_0} \left( \frac{1}{\mu_1} + \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) + \lambda_1 \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) \lambda_1
\]

It follows that any \( \frac{\lambda_1}{\mu_1} < 1 \) leads to a finite expected queue length, irrespective of \( \frac{\mu_0}{\mu_1} \).

**Step III: the conditional expected waiting time.** Finally, we can calculate \( W \) by applying Little’s Law conditioned on the state of the expert upon arrival:

\[
Q = \frac{1}{\mu_0} \pi_0 \lambda_0 + \left( \frac{1}{\mu_1} + W \right) \left( 1 - \pi_0 \right) \lambda_1 = \frac{\lambda_0}{\mu_0} \left( \frac{1}{\mu_1} + \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) + \lambda_1 \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) \lambda_1
\]

from which follows that

\[
W = \frac{\lambda_0}{\mu_0} \left( \frac{1}{\mu_1} + \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) + \lambda_1 \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) \lambda_1 = \frac{1}{\mu_1} \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) \lambda_1 = \frac{1}{\mu_1} \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right)
\]

38
Lemma 14 derives analytical expressions that will be useful later in the proofs.

**Lemma 14** The profit rate of the expert is

$$
\Pi((1, \beta_1), R, \mu) = \frac{RA \left(1 + \beta_1 A \left(\frac{1}{\mu_0} - \frac{1}{\mu_1}\right)\right) + r \frac{\Delta}{\mu_0}}{1 - \frac{\beta_1 \Delta}{\mu_1} + \frac{\Delta}{\mu_0}}
$$

The customer’s expected utility is

$$
\left\{
\begin{array}{ll}
V_0((1, \beta_1), R, \mu) = V - R - \frac{c \pi}{\mu_0} & \\
V_1((1, \beta_1), R, \mu) = V - R - \frac{c}{\mu_1} (1 - \frac{\beta_1 A}{\mu_1} (1 - \frac{\beta_1 \Delta}{\mu_1})) - \frac{c \pi}{\mu_0}
\end{array}
\right.
$$

The average queue length is

$$
Q((1, \beta_1), R, \mu) = \frac{\Lambda}{\mu_0} \frac{1}{1 - \frac{\beta_1 \Delta}{\mu_1} + \frac{\Delta}{\mu_0}} (1 - \frac{\mu_1}{\mu_0})
$$

The net arrival rate is

$$
\pi_0((1, \beta_1), \mu) = \frac{1 - \frac{\beta_1 \Delta}{\mu_1}}{1 - \frac{\beta_1 \Delta}{\mu_1} + \frac{\Delta}{\mu_0}}
$$

and the social welfare is

$$
SW_1(\beta_1, \mu) = \left(1 - \beta_1 A \left(\frac{1}{\mu_1} - \frac{1}{\mu_0}\right)\right) VA - c \left(\frac{\Lambda}{\mu_0} \frac{1}{1 - \frac{\beta_1 \Delta}{\mu_1} + \frac{\Delta}{\mu_0}} - \frac{\beta_1 \Lambda}{\mu_1} (1 - \frac{\mu_1}{\mu_0})\right)
$$

and

$$
SW_1(\beta_1, \mu) - L_1(\beta_1, \mu) = \left(V - \frac{c}{\mu_1} - c \left(\frac{1}{\mu_1} - \frac{\beta_1 \Delta}{\mu_1} + \frac{1}{\mu_0} - \frac{1}{\mu_1}\right)\right) \frac{\mu_1 A}{\mu_0}
$$

Let $\rho_0 = \frac{\Lambda}{\mu_0} \in [0, +\infty]$ and $\rho_1 = \frac{\Lambda}{\mu_1} \in [0, 1]$ and $v = \frac{VA}{c}$. Then

$$
sw_1 \geq \frac{SW_1}{c} = \frac{1 + \rho_1 \left(\frac{1}{\alpha} - 1\right)}{1 - \rho_1} v - \rho_0 \left(1 - \rho_1\right)
$$

and

$$
sw_1 - l_1 \geq \frac{SW_1 - L_1}{c} = \frac{1}{\alpha} \left(\frac{1}{\rho_0} v - 1\right) - \frac{1}{1 - \rho_1} \frac{\rho_0}{1 - \rho_1 + \rho_0}
$$

**Proof** All expressions follow immediately from Proposition 1. ■

**Proof of Lemma 2.**

The conditions for the equilibrium action in state $i = 0$ are greatly simplified as $V_0(\beta; R, \mu)$ equals $v_0(R, \mu)$ and is independent of $\beta_0$. If $v_0(R, \mu) > 0$, then, the expected utility of a randomly arriving
customer from joining an idle expert is positive, irrespective of the other customers’ strategies. So, this customer joins an idle expert with probability 1: \( \beta_0^* (R, \mu) = 1 \). Similarly if \( v_0 (R, \mu) < 0 \); \( \beta_0^* (R, \mu) = 0 \). Finally, if \( v_0 (R, \mu) = 0 \), a randomly arriving customer is indifferent between joining and not. Therefore, any randomization probability when the expert is idle is an equilibrium; \( \beta_0^* (R, \mu) = [0, 1] \). ■

**Proof of Lemma 3.**

Since our analysis is naturally restricted to \( v_0 (R, \mu) \geq 0 \), we have from Lemma 2 that \( \beta_0^* (R, \mu) = 1 \). The conditions for the equilibrium action in state \( i = 1 \); \( \beta_i^* \), are determined by \( V_i (\beta; R, \mu) \), which is a function of \( \beta_1 \). Let the solution of \( V_1 ((1, \beta_1), R, \mu) = 0 \) for \( \beta_1 \) be denoted by \( B (R, \mu) \).

Suppose \( B (R, \mu) \in (0, 1) \) and all other customers join the busy expert with probability \( B (R, \mu) \). Then, the expected utility of joining for the randomly arriving customer is zero and his best response set to \( B (R, \mu) \) is \( [0, 1] \) since he is indifferent between joining or not. Since \( B (R, \mu) \) belongs to this set, it is the unique equilibrium: \( \beta_0^* (R, \mu) = B (R, \mu) \).

Suppose \( B (R, \mu) \geq 1 \) (\( \leq 0 \)), joining with probability 1 (0) belongs to the best reaction set of a randomly arriving customer when all other customers join with probability 1 (0). Thus: \( \beta_0^* (R, \mu) = 1 \) when \( B (R, \mu) \geq 1 \) and \( \beta_0^* (R, \mu) = 0 \) when \( B (R, \mu) \leq 1 \). ■

**Proof of Proposition 4.**

For all price pairs for which \( v_0 (R, \mu) \geq 0 \) and \( v_1 (R, \mu) \leq cW_1 (0, \mu) \). As observed in Lemma 14, the profit rate \( \Pi ((1, \beta_1), R, \mu) \) for \( \beta_1 = 0 \) reduces to \( \frac{(R + \frac{r}{\mu_0})\Lambda}{1 + \frac{\Lambda}{\mu_0}} \), so, the expert’s problem is:

\[
\max_{R \in \mathbb{R}^*_+} \left( \frac{R + \frac{r}{\mu_0}}{1 + \frac{\Lambda}{\mu_0}} \right) \Lambda,
\]

s.t. \( V - \frac{c}{\mu_0} \geq \left( R + \frac{r}{\mu_0} \right) \) and \( V - \frac{c}{\mu_1} \leq cW_1 (0, \mu) \).

It is obvious that \( R + \frac{r}{\mu_0} \) needs to be as high as possible, subject to the upper bound \( V - \frac{c}{\mu_0} \). However, it may be that for no vector satisfying \( R + \frac{r}{\mu_0} = V - \frac{c}{\mu_0} \), the second inequality is satisfied. When \( r = 0 \) and \( V - \frac{c}{\mu_0} = R \), the inequality \( V - \frac{c}{\mu_1} \leq cW_1 (0, \mu) \) is always satisfied as \( \frac{1}{\mu_1} - \frac{1}{\mu_0} > 0 \) (as long as \( \frac{\beta \Lambda}{\mu_0} < 1 \)). Thus, there always exists a vector \( R \in \mathbb{R}^*_+ \) such that \( \frac{(V - \frac{c}{\mu_0})\Lambda}{1 + \frac{\Lambda}{\mu_0}} \) is the maximum profit achievable. It can be verified from Lemma 14 that this is equal to \( SW_1 (0, \mu) \). Condition (5) is nothing but the conditions \( V - \frac{c}{\mu_0} = R + \frac{r}{\mu_0} \) and \( V - \frac{c}{\mu_1} \leq cW_1 (0, \mu) \).
Proof of Proposition 5.

Overview: We fix first $\beta_1 \in [0, 1]$ and find the price vector(s) that maximize the expert’s profit rate over all price pairs $R$ that yield $(1, \beta_1)$ as equilibrium; $\{R \in \mathbb{R}_+^2 : \beta^* (R, \mu) = (1, \beta_1)\}$. In step I, we formulate the expert’s constrained optimization problem, $P(\beta_1)$ defined as $\max_{R \in \mathbb{R}_+^2} \Pi((1, \beta_1), R, \mu)$ subject to $\beta^*(R, \mu) = (1, \beta_1)$. In step II, we eliminate $R$ from the optimization problem and obtain a equivalent optimization problem, $P'(\beta_1)$ for which we observe that the objective function is increasing in $r$. $P'(\beta_1)$ has two constraints on $r$ that bind $r$ from above. In step III, we derive the expressions for the optimal price vector and the optimal profit rate in each of the two cases. In step IV, we discuss the special case $\mu_0 = \mu_1$.

Step I: The expert’s constrained optimization for given $\beta_1$: $P(\beta_1)$. The region where $\beta^*_1 (R, \mu) \in [0, 1], \Omega_1 (\mu)$ is defined by $v_0 (R, \mu) \geq 0$ and $v_1 (R, \mu) = cW_1 (\beta_1, \mu)$. As defined in Lemma 14, the profit rate $\Pi((1, \beta_1), R, \mu) = \frac{RA(1 + \beta_1 \Lambda (\frac{1}{\mu_0} - \frac{1}{\mu_1})) + r \Lambda}{1 - \frac{\Lambda}{\mu_1} + \frac{\Lambda}{\mu_0}}$, so the expert’s problem is:

$$P(\beta_1) : \max_{R \in \mathbb{R}_+^2} \frac{RA(1 + \beta_1 \Lambda (\frac{1}{\mu_0} - \frac{1}{\mu_1})) + r \Lambda}{1 - \frac{\Lambda}{\mu_1} + \frac{\Lambda}{\mu_0}}$$

s.t. $V - c \frac{1}{\mu_0} \geq \left( R + \frac{r}{\mu_0} \right)$ and $V - c \frac{1}{\mu_1} - \left( R + \frac{r}{\mu_1} \right) = cW_1 (\beta_1, \mu)$.

From the equality, we can write

$$R = V - cW_1 (\beta_1, \mu) - \frac{c + r}{\mu_1}$$

Substituting in the expert’s profit rate, we obtain that the expert’s objective function in $P(\beta_1)$ is

$$\left( V - cW_1 (\beta_1, \mu) - \frac{c}{\mu_1} \right) \Lambda_e ((1, \beta_1), \mu)
- \frac{r}{\mu_1} \Lambda_e ((1, \beta_1), \mu) + r (1 - \pi_0 ((1, \beta_1), \mu))$$

$= \left( V - cW_1 (\beta_1, \mu) - \frac{c}{\mu_1} \right) \Lambda_e ((1, \beta_1), \mu) - r \Lambda_0 ((1, \beta_1), \mu) \left( \frac{1}{\mu_1} - \frac{1}{\mu_0} \right)
- r \left( \frac{1}{\mu_1} \Lambda_1 (1, \beta_1, \mu) + \frac{1}{\mu_0} \Lambda_0 (1, \beta_1, \mu) \right) + r (1 - \pi_0 ((1, \beta_1), \mu))$

$= \left( V - cW_1 (\beta_1, \mu) - \frac{c}{\mu_1} \right) \Lambda_e ((1, \beta_1), \mu) + r \Lambda_0 ((1, \beta_1), \mu) \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right)$

The last step is because the fraction of time that the expert is busy $(1 - \pi_0 ((1, \beta_1), \mu))$ is determined by the sum of the arrival rate when the expert is idle times the exponential service rate when the expert is idle.
and the arrival rate when the expert is busy times the exponential service time when the expert is busy \( (\frac{1}{\mu_1} \Lambda_1 ((1, \beta_1), \mu)) \). We observe that the objective function is increasing in \( r \) as \( \frac{1}{\mu_0} - \frac{1}{\mu_1} > 0 \) and the optimization problem is solved by setting \( r \) as high as possible.

**Step II: The expert’s equivalent constrained optimization for given \( \beta_1 \); \( P' (\beta_1) \).**

We can rewrite the objective function in terms of the social welfare as follows:

\[
V \Lambda_e ((1, \beta_1), \mu) - c W_1 (\beta_1, \mu) - \Lambda_0 ((1, \beta_1), \mu) + (1 - c) \frac{1}{\mu_1} \Lambda_0 ((1, \beta_1), \mu)
\]

The last step is because the expected number of customers in the system, \( Q_1 (\beta_1, \mu) \), is equal to \( \frac{1}{\mu_0} \Lambda_0 ((1, \beta_1), \mu) + (1 - c) \frac{1}{\mu_1} \Lambda_0 ((1, \beta_1), \mu) \).

Substituting \( R \) from Equation 21 in the first constraint of \( P (\beta_1) \), we obtain:

\[
V - c \frac{1}{\mu_0} \geq V - c \frac{1}{\mu_1} - c W_1 (\beta_1, \mu) + r \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right)
\]

or

\[
r_l (\beta_1, \mu) = c \left( W_1 (\beta_1, \mu) - 1 \right) \geq r \Leftrightarrow \Delta (r, \beta_1, \mu) \geq 0 \tag{22}
\]

and, as \( R \in \mathbb{R}^2_{+} \), or \( R \geq 0 \), we obtain from Equation 21 that

\[
r_{II} (\beta_1, \mu) = \mu_1 \left( V - c \left( \frac{1}{\mu_1} + W_1 (\beta_1, \mu) \right) \right) \geq r. \tag{23}
\]

We obtain thus two upper bounds on \( r \). The expert’s equivalent optimization problem is

\[
P' (\beta_1) : \max_{r \in \mathbb{R}^2_{+}} SW_1 (\beta_1, \mu) - \Lambda_0 ((1, \beta_1), \mu) \Delta (r, \beta_1, \mu)
\]

s.t. \( r_l (\beta_1, \mu) \geq r \) and \( r_{II} (\beta_1, \mu) \geq r \)

**Step III: The expert’s optimal profit rate for given \( \beta_1 \).**
Substitute for the highest $r$ for which Equation 23, is satisfied, $r_{II} (\beta_1, \mu)$, in $\Delta (r, \beta_1, \mu)$:

$$
\Delta_0 (\beta_1, \mu) \triangleq \Delta (r_{II} (\beta_1, \mu), \beta_1, \mu) = cW_1 (\beta_1, \mu) - \left( \frac{1}{\mu_1} \left( V - c \left( \frac{1}{\mu_1} + W_1 (\beta_1, \mu) \right) \right) + c \left( \frac{1}{\mu_0} - \frac{1}{\mu_1} \right) \right)
$$

$$
= V \left( 1 - \frac{\mu_1}{\mu_0} \right) + cW_1 (\beta_1, \mu) \frac{\mu_1}{\mu_0}
$$

If $\Delta_0 (\beta_1, \mu) \geq 0$, this means that when Equation 23 is binding (i.e. $R = 0$), Equation 22 is not binding. Then, the optimal profit rate is determined by:

$$
\Pi_1 (\beta_1, \mu) = SW_1 (\beta_1, \mu) - \Lambda_0 (r_{II}, \mu) \Delta_0 (\beta_1, \mu)
$$

and the optimal price vector is thus $r_{II} (\beta_1, \mu)$ for the billing rate and $R_{II} (\beta_1, \mu) = 0$ for the fixed fee.

Otherwise, if $\Delta_0 (\beta_1, \mu) < 0$, Equation 22 is binding and Equation 23 is not binding and

$$
\Pi_1 (\beta_1, \mu) = SW_1 (\beta_1, \mu)
$$

The optimal billing rate is thus $r_I (\beta_1, \mu)$ and fixed fee can be obtained by substituting $r_I$ in Equation 21, which yields

$$
R_I (\beta_1, \mu) = V - c \frac{\mu_1}{\mu_1 - \mu_0} W_1 (\beta_1, \mu)
$$

$$
= - \frac{\mu_0}{\mu_1 - \mu_0} \Delta_0 (\beta_1, \mu)
$$

Note that $\Delta_0 (\beta_1, \mu)$ is monotone increasing $\beta_1$ (as $W_1$ increases in $\beta_1$ by Proposition 1). Let $\overline{B} (\mu)$ solve $\Delta_0 (\beta_1, \mu) = 0$. If $\overline{B} (\mu) \in [0, 1]$, then, for $\beta_1 \in [0, \overline{B} (\mu)]$, $\Delta_0 (\beta_1, \mu) < 0$ and thus $\Pi_1 (\beta_1, \mu)$ is determined by Equation 25, otherwise, $\Pi_1 (\beta_1, \mu)$ is determined by Equation 24.

**Step IV:** When $\mu_0 = \mu_1$, it is obvious that

$$
\Delta_0 (\beta_1, \mu) = cW_1 (\beta_1, \mu) > 0
$$

and therefore, only case I is applicable and Equation 24 determines the optimal profits. Furthermore, $\Delta (r, \beta_1, \mu)$ is independent of $r$. Thus, there exist multiple price pairs $R$ that yield the same profit rate for the expert; all are determined by Equation 21. ■

**Proof of Proposition 6.**

We assume that $\Omega_2 (\mu)$ is non-empty, or, $v_1 (R, \mu) \geq cW_1 (1, \mu)$, otherwise, the price optimization in $\Omega_2 (\mu)$ is irrelevant. The conditions defining $\Omega_2 (\mu)$ are $v_0 (R, \mu) \geq 0$ and $v_1 (R, \mu) \geq cW_1 (1, \mu)$. Substi-
tuting $\beta_0 = \beta_1 = 1$ in to the the profit rate (derived in from Lemma 14) gives $\frac{RA(1+\Lambda(\frac{1}{\mu_0} - \frac{1}{\mu_1})) + r\Lambda}{1-\frac{\Lambda}{\mu_1} + \frac{\Lambda}{\mu_0}}$. The expert’s optimization problem is

$$\max_{R \in \mathbb{R}^2} \left\{ \frac{R\Lambda(1 + \Lambda(\frac{1}{\mu_0} - \frac{1}{\mu_1})) + r\Lambda}{1 - \frac{\Lambda}{\mu_1} + \frac{\Lambda}{\mu_0}} \right\},$$

s.t. $V - c \frac{1}{\mu_0} \geq \left( R + \frac{r}{\mu_0} \right)$ and $V - c \frac{1}{\mu_1} - cW_1(1, \mu) \geq \left( R + \frac{r}{\mu_1} \right)$.

This objective function increases both in $R$ and $r$. As a result, at least one of the constraints must be binding. Assume that $\frac{c}{\mu} = V - \frac{c}{\mu_0} - R$, then, the objective function becomes:

$$\frac{R\Lambda^2(\frac{1}{\mu_0} - \frac{1}{\mu_1}) + \Lambda(V - \frac{c}{\mu_0})}{1 - \frac{\Lambda}{\mu_1} + \frac{\Lambda}{\mu_0}}$$

which is increasing in $R$ as $\frac{1}{\mu_0} - \frac{1}{\mu_1} > 0$. The second constraint evaluated at $\frac{c}{\mu} = V - \frac{c}{\mu_0} - R$, is $V - \frac{c}{\mu_1} - cW_1(1, \mu) \geq R + \frac{\mu_0}{\mu_1} \frac{r}{\mu_0} = \left( 1 - \frac{\mu_0}{\mu_1} \right) R + \frac{\mu_0}{\mu_1} \left( V - \frac{c}{\mu_0} \right)$. Thus, the maximum profits are obtained at $R = \max \left\{ 0, \frac{V - cW_1(1, \mu) - \frac{c}{\mu_0}(V - c)}{(1 - \frac{\mu_0}{\mu_1})} \right\}$: if the two constraints do not intersect in $\mathbb{R}_+^2$ and non-negative otherwise. The solution for $r$ is always positive: If $R = 0$, then, $\frac{c}{\mu_0} = V - \frac{c}{\mu_0} > 0$, or if $R > 0$, then

$$\frac{r}{\mu_0} = c \frac{1}{\mu_1} - \frac{1}{\mu_0} + \frac{W_1(1, \mu)}{1 - \frac{\mu_0}{\mu_1}} = c \frac{\frac{1}{\mu_1} - \frac{1}{\mu_0}}{1 - \frac{\mu_0}{\mu_1}} > 0.$$ 

\[ \blacksquare \]

**Proof of Proposition 7.**

**Overview:** In order to reduce the notational burden, we prove the unimodality of $SW_1$ and $SW_1 - L_1$ in $\beta_1$ by defining $sw_1 = \frac{SW_1}{c}$ and $sw_1 - l_1 = \frac{SW_1 - L_1}{c}$ as functions of $\rho_1 = \frac{\beta_1 \Lambda}{\mu_1}$ (from Lemma 14) and prove the unimodality of $sw_1$ and $sw_1 - l_1$ in $\rho_1$. In Equation 7, we have reduced the problem to a maximization problem of unimodal functions over separate, but adjacent intervals; $[0, p]$ and $[p, 1]$, where $p = \frac{\mu(\mu)\Lambda}{\mu_1}$. We check unimodality for both profit functions over $[0, 1]$ and then check the derivatives at the common boundary $p$. We determine the first and second order derivative and evaluate the second order derivative at a value that makes the first order derivative equal to zero. We find that at that point the second order derivative is negative, and therefore, the function is unimodal. This is done in steps I and II. In step III, we prove that the derivatives of $sw_1$ and $sw_1 - l_1$ at their intersection point (i.e. where $l_1 = 0$ and $\rho_1 = p$) can be ranked; $\frac{dsw_1}{d\rho_1} \bigg|_{l_1=0} > \frac{d(sw_1-l_1)}{d\rho_1} \bigg|_{l_1=0}$. An implication of this is that only one local maximum needs to be
Figure 7: $\Pi_1(\beta_1)$ for $c = 0.075$ and $\Lambda = 0.9$ (left) $c = 0.175$ and $\Lambda = 0.61$ (middle), and $c = 0.2$ and $\Lambda = 0.35$ (right) with $\mu_0 = 0.65$, $\mu_1 = 1$ and $V = 1$. The interior maximum is $\beta_1^I$ (type I) $\overline{B}$ (type III) and $\beta_1^{II}$ (type II), respectively.

considered, or a boundary of the two interval. The three possible situations (an interior solution over $[0, \overline{\rho}]$, $\overline{\rho}$ and an interior solution over $[\overline{\rho}, 1]$) are illustrated in Figure 7.

**Step I:** $sw_1$ is unimodal in $\rho_1 \in [0, 1]$: Recall from Lemma 14 that

$$sw_1 = \frac{1 + \rho_1 \left( \frac{1}{\alpha} - 1 \right)}{1 - \rho_1 + \rho_0} - \frac{\rho_0 \rho_1 \left( \frac{1}{\alpha} - 1 \right)}{1 - \rho_1 + \rho_0} - \frac{\rho_0 \rho_1 \left( \frac{1}{\alpha} - 1 \right)}{1 - \rho_1 + \rho_0}$$

We determine the first and second order derivatives:

$$\frac{dA}{d\rho_1} = 1 + \left( \frac{1}{\alpha} - 1 \right) \left( 1 + \rho_0 \right) \left( \rho_1 - \rho_0 - 1 \right)^2$$

$$\frac{dB}{d\rho_1} = \frac{\rho_0 (\rho_0 - 2 \rho_1 + 2)}{(\rho_1 - 1)^2 (\rho_0 - \rho_1 + 1)^2}$$

and

$$\frac{dC}{d\rho_1} = \frac{\rho_0 \left( \frac{1}{\alpha} - 1 \right) (\rho_0 + 1)}{(\rho_1 - \rho_0 - 1)^2}$$

and

$$\frac{d^2 A}{d\rho_1^2} = 2 \left( 1 + \left( \frac{1}{\alpha} - 1 \right) (1 + \rho_0) \right) \left( 1 - \rho_1 + \rho_0 \right)^3$$

$$\frac{d^2 B}{d\rho_1^2} = 2 \rho_0 \left( 3 \rho_0 - 6 \rho_1 - 3 \rho_1 \rho_0 + 3 \rho_1^2 + \rho_0^2 + 3 \right) \left( 1 - \rho_1 \right)^4 (1 - \rho_1 + \rho_0)^3$$

and

$$\frac{d^2 C}{d\rho_1^2} = 2 \rho_0 \left( \frac{1}{\alpha} - 1 \right) (\rho_0 + 1) \left( 1 - \rho_1 + \rho_0 \right)^3 > 0$$

45
Combining these terms, we can write \( \frac{d^2 sw_1}{dp_1^2} \) as a function of \( \frac{dsw}{dp_1} \):

\[
\frac{dsw}{dp_1} = \frac{dA}{dp_1} - \frac{d(B + C)}{dp_1} = v \left( \left( \frac{1}{\alpha} - 1 \right) (\rho_0 + 1) + 1 \right) - \rho_0 \left( \frac{\rho - 2\rho_1 + 2}{(\rho_1 - 1)^2} + \left( \frac{1}{\alpha} - 1 \right) (\rho_0 + 1) \right)
\]

\[
\frac{d^2 sw_1}{dp_1^2} = \frac{d^2 A}{dp_1^2} - \frac{d^2 (B + C)}{dp_1^2} = 2 v \left( \left( \frac{1}{\alpha} - 1 \right) (\rho_0 + 1) + 1 \right) - \rho_0 \left( \frac{\rho - 2\rho_1 + 2}{(\rho_1 - 1)^2} + \left( \frac{1}{\alpha} - 1 \right) (\rho_0 + 1) \right)
\]

\[
= 2 \frac{dsw}{dp_1} - \frac{\rho_0}{(1-\rho_1)^2 (1-\rho_1 + \rho_0)}
\]

Note that the FOC \( \left( \frac{dsw}{dp_1} = 0 \right) \) has the following structure: \( \Psi = \frac{2(1-\rho_1) + \phi}{(1-\rho_1)^2} \Rightarrow \rho_1 = 1 - \frac{1}{\Psi} \left( 1 \pm \sqrt{\phi \Psi + 1} \right) \), where \( \phi \) and \( \Psi \) can be identified and \( \Psi > 0 \). The negative root can be discarded, \( 1 - \frac{1}{\Psi} (1 - \sqrt{\phi \Psi + 1}) > 1 \), for which the system is not stable (see Proposition 1 which is valid only if \( \lambda_1 < \mu_1 \) or \( \rho_1 < 1 \)). Therefore, only the positive root needs to be considered as \( 1 - \frac{1}{\Psi} (1 + \sqrt{\phi \Psi + 1}) < 1 \). It follows that when \( \frac{dsw}{dp_1} = 0 \), we obtain that

\[
\frac{d^2 sw_1}{dp_1^2} = -2 \frac{\rho_0}{(1-\rho_1)^2 (1-\rho_1 + \rho_0)} < 0
\]

for \( \rho_1 \in (0, 1) \). Thus, \( sw_1 \) is unimodal in \( \rho_1 \) over \([0, 1]\).

**Step II: \( sw_1 - l_1 \) is unimodal in \( \rho_1 \in [0, 1] \):**

Recall from Lemma 14 that

\[
sw_1 - l_1 = \left( \frac{1}{\alpha} \left( v \frac{1}{\rho_0} - 1 \right) - \frac{1}{1-\rho_1} \right) \frac{\rho_0}{1-\rho_1 + \rho_0}
\]

We determine the first and second order derivatives:

\[
\frac{d (sw_1 - l_1)}{dp_1} = \rho_0 \left( \frac{\rho - 2\rho_1 + 2}{(\rho_1 - 1)^2} + \frac{1}{\alpha} \left( v \frac{1}{\rho_0} - 1 \right) \right) \frac{\rho_0}{(\rho_0 - \rho_1 + 1)^2}
\]

\[
\frac{d^2 (sw_1 - l_1)}{dp_1^2} = -2 \frac{\rho_0 \left( \frac{\rho - 2\rho_1 + 2}{(\rho_1 - 1)^2} + \frac{1}{\alpha} \left( v \frac{1}{\rho_0} - 1 \right) \right) \rho_0}{(\rho_0 - \rho_1 + 1)^2}
\]

Combining these terms, we can write \( \frac{d^2 (sw_1 - l_1)}{dp_1^2} \) as a function of \( \frac{d(sw_1 - l_1)}{dp_1} \):

\[
\frac{d^2 (sw_1 - l_1)}{dp_1^2} = -2 \frac{d(sw_1 - l_1)}{dp_1} + \frac{2\rho_0}{(1-\rho_1)^2}
\]

Note that the FOC \( \left( \frac{d(sw_1 - l_1)}{dp_1} = 0 \right) \) has the following structure: \( \Psi = \frac{2(1-\rho_1) + \phi}{(1-\rho_1)^2} \Rightarrow \rho_1 = 1 - \frac{1}{\Psi} \left( 1 \pm \sqrt{\phi \Psi + 1} \right) \), where \( \phi \) and \( \Psi \) can be identified and \( \Psi > 0 \). The negative root can be discarded, as we assume that \( \rho_1 < 1 \).
Therefore, only the positive root needs to be considered. It follows that when \( \frac{d(sw_1 - l_1)}{d\rho_1} = 0 \),
\[
\frac{d^2 (sw_1 - l_1)}{d\rho_1^2} = -\frac{2\rho_0}{(1 - \rho_1 + \rho_0)(1 - \rho_1)^2} < 0
\]
when \( \rho_1 < 1 \). Thus, \( (sw_1 - l_1) \) is unimodal in \( \rho_1 \) over \([0, 1]\).

**Step III:** Now, we prove that when \( l_1 = 0 \), or alternatively, when \( \rho = \frac{\Lambda}{\mu} \), that
\[
\frac{ds_{sw_1}}{d\rho_1} \bigg|_{l_1=0} > \frac{d(sw_1 - l_1)}{d\rho_1} \bigg|_{l_1=0}
\]
which implies that the maximum will either be attained at 0, \( \rho \), or the local maximum of \( sw_1 \) on \([0, \rho] \) or the local maximum of \( sw_1 - l_1 \) on \([\rho, 1]\). The structure of the optimization problem is the following:
\[
\max_{\rho_1 \in (0, 1)} \begin{cases} 
sw_1 & \rho_1 \in [0, \rho] \\
sw_1 - l_1 & \rho_1 \in [\rho, 1]
\end{cases}
\]
As both functions are unimodal, they either have a local maximum, or, always increase, or, always decrease in their respective intervals. We claim that when \( sw_1 \) has a local maximum in \([0, \rho] \), \( sw_1 - l_1 \) cannot have a local maximum in \([\rho, 1]\) and vice versa. Assume that \( sw_1 \) has a local maximum in \([0, \rho] \). Then, \( sw_1 \) must be decreasing at \( \rho \), and if Equation 26 holds, \( sw_1 - l_1 \) must be decreasing at \( \rho \) as well and cannot attain a local maximum in \([\rho, 1]\). Similarly, assume that \( sw_1 - l_1 \) has a local maximum in \([\rho, 1]\). Then, \( sw_1 - l_1 \) must be increasing at \( \rho \), and if Equation 26 holds, \( sw_1 \) must be increasing at \( \rho \) as well and cannot attain a local maximum in \([0, \rho] \). We prove Equation 26 as follows: We rewrite:
\[
sw_1 = \left(1 + \rho_1 \left(\frac{1}{\alpha} - 1\right)\right) v - \rho_0 \left(\frac{v}{\rho_0} + \rho_1 \left(\frac{1}{\alpha} - 1\right)\right)
\]
\[
\frac{ds_{sw}}{d\rho_1} = \frac{sw_1}{1 - \rho_1 + \rho_0} - \rho_0 \left(\rho_1 \left(1 - \alpha\right) \left(\rho_1 - 2\right) + v \left(\rho_1 - 1\right)^2 \left(\alpha - 1\right)\right) / \alpha \left(\rho_1 - 1\right)^2 \left(\rho_1 - \rho_0 - 1\right)
\]
\[
\frac{d(sw_1 - l_1)}{d\rho_1} = \left(\frac{1}{\alpha} \left(\frac{v}{\rho_0} - 1\right) - \frac{1}{1 - \rho_1}\right) \frac{\rho_0}{1 - \rho_1 + \rho_0}
\]
\[
\frac{1}{\rho_1 - \rho_0 - 1} \left(s_{sw_1} - l_1\right) + \frac{sw_1 - l_1}{1 - \rho_1 + \rho_0} + \frac{\rho_0}{\left(\rho_1 - 1\right)^2 \left(\rho_1 - \rho_0 - 1\right)}
\]
At \( \rho_1 = \rho \), we have \( sw_1 - l_1 = sw_1 \). Then
\[
\frac{ds_{sw_1}}{d\rho_1} \bigg|_{l_1=0} - \frac{d(sw_1 - l_1)}{d\rho_1} \bigg|_{l_1=0} = v \left(\rho_1 - 1\right)^2 \frac{1 - \alpha}{\alpha} - \rho_0 \left(\rho_1 \left(1 - \alpha\right) \left(2 - \rho_1\right)\right) + \rho_0
\]
\[
= \frac{1}{\alpha} \left(1 - \alpha\right) \left(v - \rho_0\right) / \left(1 - \rho_1 + \rho_0\right) > 0
\]
which implies Equation 26. ■

**Proof of Proposition 8.** The calculation of the derivatives is tedious, but, straightforward. Recall from Lemma 14 that

\[
sw_1 = \frac{1 + \frac{\Lambda}{\mu_1} \left( \frac{\mu_1}{\mu_0} - 1 \right)}{1 - \frac{\beta_1}{\mu_1} + \frac{\Lambda}{\mu_0}} \mu_1 - \frac{\Lambda}{\mu_0} \left( \frac{1}{\mu_1} \right) + \frac{\beta_1}{\mu_1} \left( \frac{\mu_1}{\mu_0} - 1 \right)
\]

and

\[
sw_1 - l_1 = \frac{\mu_1}{\mu_0} \left( 1 - \frac{\beta_1}{\mu_1} + \frac{\Lambda}{\mu_0} \right) \mu_1 - \frac{\mu_1}{\mu_0} \left( 1 - \frac{\beta_1}{\mu_1} + \frac{\Lambda}{\mu_0} \right) \mu_1 - \frac{1}{\mu_0} \left( 1 - \frac{\beta_1}{\mu_1} + \frac{\Lambda}{\mu_0} \right)
\]

(i-a) \( \frac{\partial sw_1}{\partial \mu_0} \mid_{\beta_1=0} \): When \( \beta_1 = 0 \), then \( \mu_1 = 0 \) and \( \frac{\partial v - \Lambda}{\partial \mu_0} = \Lambda(v+1)(\Lambda+\mu_0)^2 > 0 \)

(i-b) \( \frac{\partial sw_1}{\partial \mu_1} \mid_{\beta_1=0} \): When \( \beta_1 = 0 \), then \( \mu_1 = 0 \) and \( \frac{\partial sw_1}{\partial \mu_1} = 0 \)

(ii-a) \( \frac{\partial sw_1}{\partial \mu_0} \): \[
\frac{\partial A}{\partial \mu_0} = \frac{1}{\mu_0} \left( 1 - \beta_1 \right) \left( 1 - \frac{\Lambda}{\mu_1} \right) \Lambda > 0
\]

\[
\frac{\partial B}{\partial \mu_0} = -\frac{\Lambda \mu_0}{\mu_1} \left( 1 - \frac{\Lambda}{\mu_1} \right) \left( 2 - \frac{\mu_0}{\mu_1} \right) + \frac{\mu_0}{\mu_1} \left( \frac{\Lambda}{\mu_1} - 1 \right) < 0
\]

As all terms of \( \frac{\partial A}{\partial \mu_0} \) are positive, \( \frac{\partial sw_1}{\partial \mu_0} > 0 \).

(ii-b) \( \frac{\partial sw_1}{\partial \mu_1} \): \[
\frac{\partial A}{\partial \mu_1} = \frac{1}{\mu_1} \left( 1 - \beta_1 \right) \Lambda \left( 1 + \Lambda - \frac{\Lambda}{\mu_1} \right) > 0
\]

\[
\frac{\partial B}{\partial \mu_1} = -\frac{\Lambda}{\mu_1} \left( 1 + \frac{\mu_0}{\mu_1} - \frac{\Lambda}{\mu_1} \right) \left( 1 - \frac{\Lambda}{\mu_1} \right) \left( 1 - \frac{\Lambda}{\mu_1} \right) < 0
\]

As all terms of \( \frac{\partial A}{\partial \mu_1} \) are positive, \( \frac{\partial sw_1}{\partial \mu_1} > 0 \).

(iii-a) \( \frac{\partial (sw_1 - l_1)}{\partial \mu_0} \): \[
\frac{\partial A}{\partial \mu_0} = \frac{\mu_1}{\mu_0} \left( \frac{\Lambda}{\mu_1} - 1 \right) \Lambda \left( \frac{\Lambda}{\mu_1} - 1 \right) > 0
\]

\[
\frac{\partial B}{\partial \mu_0} = \frac{\Lambda}{\mu_1} \mu_1 \left( 2 - \frac{\Lambda}{\mu_1} \right) \left( 1 - \frac{\Lambda}{\mu_1} \right) < 0
\]

\[
\frac{\partial C}{\partial \mu_0} = -\frac{\Lambda}{\mu_1} \left( \frac{\Lambda}{\mu_1} - 1 \right) < 0
\]

48
It follows that

\[
\frac{\partial A}{\partial \mu_0} v - \frac{\partial B}{\partial \mu_0} - \frac{\partial C}{\partial \mu_0} = \frac{\mu_1}{\mu_0} \frac{1}{1 - \frac{\beta_1}{\mu_1}} v + \frac{\mu_1}{\mu_0} \frac{1 - \frac{\beta_1}{\mu_1}}{1 - \frac{\beta_1}{\mu_1}} (2 \left( 1 - \frac{\beta_1}{\mu_1} \right) + \frac{\Lambda}{\mu_0})
\]

and thus as the first term is negative and the second term is positive, we have that \(\frac{\partial (sw_1 - l_1)}{\partial \mu_0} \leq 0\); negative for large values of \(v\), positive for small values of \(v\).

(iii-b) \(\frac{\partial (sw_1 - l_1)}{\partial \mu_1}\):

\[
sw_1 - l_1 = \mu_1 \frac{1}{\mu_0} \frac{\Lambda}{\mu_0} \frac{1}{1 - \frac{\beta_1}{\mu_1}} v + \mu_1 \frac{1}{\mu_0} \frac{1 - \frac{\beta_1}{\mu_1}}{1 - \frac{\beta_1}{\mu_1}} (2 \left( 1 - \frac{\beta_1}{\mu_1} \right) + \frac{\Lambda}{\mu_0})
\]

\[
\frac{\partial A}{\partial \mu_1} = \frac{\left( \frac{\mu_1}{\mu_0} - \beta_1 \right) \Lambda}{\mu_1} + \frac{1 - \frac{\beta_1}{\mu_1}}{1 - \frac{\beta_1}{\mu_1}} (2 \left( 1 - \frac{\beta_1}{\mu_1} \right) + \frac{\Lambda}{\mu_0})
\]

\[
\frac{\partial B}{\partial \mu_1} = \frac{\left( \frac{\mu_1}{\mu_0} - \beta_1 \right) \Lambda}{\mu_1} + \frac{1 - \frac{\beta_1}{\mu_1}}{1 - \frac{\beta_1}{\mu_1}} (2 \left( 1 - \frac{\beta_1}{\mu_1} \right) + \frac{\Lambda}{\mu_0})
\]

\[
\frac{\partial C}{\partial \mu_1} = -\frac{\Lambda}{\mu_0} \frac{1}{1 - \frac{\beta_1}{\mu_1}} - \frac{2 \left( 1 - \frac{\beta_1}{\mu_1} \right) + \frac{\Lambda}{\mu_0}}{1 - \frac{\beta_1}{\mu_1}}
\]

It follows that

\[
\frac{\partial A}{\partial \mu_1} v - \frac{\partial B}{\partial \mu_1} = \left( v - \frac{\Lambda}{\mu_0} \right) \frac{1}{\mu_0} \frac{\mu_1}{\mu_0} \frac{1 - \frac{\beta_1}{\mu_1}}{1 - \frac{\beta_1}{\mu_1}} > 0
\]

and thus as \(v = \frac{\nu \Lambda}{\epsilon} > \frac{\Lambda}{\mu_0}\) (Assumption (1)) and \(\frac{\mu_1}{\mu_0} > 1 > \beta_1\), all terms of \(\frac{\partial A}{\partial \mu_0} v - \frac{\partial B}{\partial \mu_0} - \frac{\partial C}{\partial \mu_0}\) are positive, or \(\frac{\partial (sw_1 - l_1)}{\partial \mu_1} > 0\).

**Proof of Corollaries 9 and 10.**

(i) Corollary 9: \(\frac{\partial \Pi(\mu)}{\partial \mu_1} \frac{\partial \Pi(\mu)}{\partial \mu_0} \leq 0\) follows immediately from Proposition 8(iii).

(ii) Corollary 10: \(\frac{\partial \Pi(\mu)}{\partial \mu_1} \geq 0\) can be proven as follows:

1. When the solution \(\beta^*_1(\mu)\) is 0 or 1, the Proposition is proven immediately with Proposition 8(i) and (ii).
2. When the solution $\beta_1^* (\mu)$ is an interior maximizer, the Proposition is proven immediately with the envelope theorem:

\[
\cdot \frac{\partial \Pi (\mu)}{\partial \mu_1} = \frac{\partial (SW_1 (\beta_1, \mu) - L_1 (\beta_1, \mu))}{\partial \beta_1} \frac{\partial \beta_1^* (\mu)}{\partial \mu_1} = \frac{\partial (SW_1 (\beta_1, \mu) - L_1 (\beta_1, \mu))}{\partial \mu_1} > 0
\]

with Proposition 8(ii).

\[
\cdot \frac{\partial \Pi (\mu)}{\partial \mu_1} = \frac{\partial SW_1 (\beta_1, \mu)}{\partial \beta_1} \frac{\partial \beta_1^* (\mu)}{\partial \mu_1} = \frac{\partial SW_1 (\beta_1, \mu)}{\partial \mu_1} > 0 \text{ with Proposition 8(ii).}
\]

3. When the solution $\beta_1^* (\mu)$ is the corner $\overline{B} (\mu)$, then let $\mu' = (\mu_0, \mu'_1)$ with $\mu'_1 > \mu_1$. With Proposition 8(ii), $SW_1 (\overline{B} (\mu), \mu') \geq SW_1 (\overline{B} (\mu), \mu) = \hat{\Pi} (\mu)$. As $\overline{B} (\mu)$ belongs to the optimization set of $\hat{\Pi} (\mu')$, it follows that $\hat{\Pi} (\mu') \geq SW_1 (\overline{B} (\mu), \mu')$. Thus, we have obtained that $\hat{\Pi} (\mu') \geq \hat{\Pi} (\mu)$.

**Proof of Proposition 11:**

**Overview:** From Figure 4, we know that regions a, b, c, f only are relevant for $c$ small. In order to determine an approximation of the border of regions e and f, where $\beta_1^f (\mu_a) = 1$ and of the border of regions a and b, where $\beta_1^f (\mu_h) = 1$, we first develop analytical expressions in Step I, then, approximate these expressions when $c \to 0^+$ in Step II and finally solve for $\beta_1^f (\mu_a) = 1$ and $\beta_1^f (\mu_h) = 1$ in Step III.

**Step I: Characterization of $\beta_1^f (\mu_a)$ and $\beta_1^f (\mu_h)$.** We will work with $\rho_1^I (\rho_1^I)$ instead of $\beta_1^f (\beta_1^f)$. From Proposition 8, we can write

\[
sw_1^I = \frac{(1 + \rho_1^I (\frac{1}{\alpha} - 1)) v - \rho_1^I \left(\frac{1}{\alpha - 1} + \rho_1^I \left(\frac{1}{\alpha} - 1\right)\right)}{1 - \rho_1^I + \rho_0^I}
\]

as a function of $\rho_1^I = \frac{\beta_1^f \mu_a}{\rho_1}$ and $\rho_0^I = \frac{\mu}{\rho_1}$, where $\rho_1^I$ is determined by the equality $\frac{sw_1^I}{\rho_1} = 0$, or

\[
\left(\frac{1}{\alpha} - 1\right) (\rho_0^I + 1) + 1) v - \rho_0^I \left(\frac{\rho_0^I - 2 \rho_1^I + 2}{(1 - \rho_1^I)^2} + \left(\frac{1}{\alpha} - 1\right) (\rho_0^I + 1)\right) = 0.
\]

For honest service, we have that $\rho_0^I = \frac{\mu}{\rho_1}$, and $\rho_1^I = \frac{\beta_1^f \mu_a}{\rho_1}$, so we can write

\[
sw_1^I - t_1^I = \frac{v \frac{1}{\rho_1} - 1 - \frac{1}{1 - \rho_1^I}}{1 - \rho_1^I + \rho_0^I},
\]

where $\rho_1^I$ is determined by the equality $\frac{sw_1^I - t_1^I}{\rho_1^I} = 0$, or

\[
\frac{2 \rho_1^I - \rho_0^I - \frac{2}{(1 - \rho_1^I)^2}}{\rho_0^I} + \left(\frac{1}{\rho_0^I} - 1\right) = 0.
\]
Both FOCs in Equations 28 and 30 can be written as:

\[
\frac{(\frac{1}{\alpha} - 1)(\rho_0 + 1) + 1}{\rho_0} v - \frac{(\frac{1}{\alpha} - 1)(\rho_0 + 1)}{\rho_0} = \frac{2 - 2\rho'_I + \frac{\phi_I}{\rho_0}}{(1 - \rho'_I)^2} \quad \text{and} \quad v \frac{1}{\alpha \rho_0} - 1 = \frac{2 - 2\rho''_I + \frac{\phi_{II}}{\rho_0}}{(1 - \rho''_I)^2}
\]

As a result, both FOCs have a similar structure:

\[
\Psi = \frac{2(1 - \rho) + \phi}{(1 - \rho)^2}
\]

where

\[
\Psi_I = \frac{(\frac{1}{\alpha} - 1)(\rho_0 + 1) + 1}{\rho_0} v - \frac{(\frac{1}{\alpha} - 1)(\rho_0 + 1)}{\rho_0} \quad \text{for} \quad I
\]

\[
\Psi_{II} = v \frac{1}{\alpha \rho_0} - 1 \quad \text{and} \quad \phi_{II} = \alpha \rho_0.
\]

Solving Equation 31 for \( \rho \), we obtain that \( \rho_{I,2}^{\text{opt}} = 1 - \frac{1}{\Psi} (1 + \sqrt{\Psi}) \). As the net arrival rate needs to be less than the service rate in order to have a stable queue, the root \( 1 - \frac{1}{\Psi} (1 - \sqrt{\Psi}) \) can be excluded. Thus, the unique point that satisfies the FOC is \( \rho^{\text{opt}} = 1 - \frac{1}{\Psi} (1 + \sqrt{\Psi}) \). Substituting \( \Psi_I, \Psi_{II}, \phi_I \) and \( \phi_{II} \) back into \( \rho^{\text{opt}} \), we obtain that

\[
\rho_I = 1 - \frac{1}{\Psi_I} \left( 1 + \sqrt{\phi_I \Psi_I + 1} \right) \quad \text{and} \quad \rho_{II} = 1 - \frac{1}{\Psi_{II}} \left( 1 + \sqrt{\phi_{II} \Psi_{II} + 1} \right)
\]

\[
\Psi_I = \left( \frac{1 - 1}{\alpha} \left( \frac{\Lambda}{\mu_0} + 1 \right) + 1 \right) \frac{\mu_0 V}{\alpha c} - \left( \frac{1 - 1}{\alpha} \left( \frac{\Lambda}{\mu_0} + 1 \right) \right), \phi_I = \frac{\Lambda}{\mu_0}
\]

\[
\Psi_{II} = \frac{V}{\alpha c} \left( \frac{\mu_0}{\alpha} - 1, \phi_{II} = \frac{\alpha}{\mu_0} \Lambda \right)
\]

**Step II: Limits of \( \beta_I^I (\mu_a) \) and \( \beta_{II}^I (\mu_h) \) for \( c \to 0^+ \)**: Changing the variable \( c \) into \( \gamma \) satisfying \( \gamma^2 = \frac{\gamma^2}{\mu_1} \), and introducing \( \tau^2 = \left( \frac{1}{\mu} - 1 \right) \left( \frac{\xi^2}{\mu} + 1 \right) + 1 \) and \( \xi^2 = \frac{\Lambda}{\mu_1} \), we obtain that

\[
\Psi_I = \frac{\tau^2 \alpha}{\gamma^2} - \left( \frac{1 - 1}{\alpha} \left( \frac{\Lambda}{\mu_0} + 1 \right) \right), \phi_I = \frac{\xi^2}{\alpha}
\]

\[
\Psi_{II} = \frac{1}{\gamma^2} - 1, \phi_{II} = \xi^2
\]

and the optimal \( \rho_I \) and \( \rho_{II} \) are determined as follows:

\[
\rho_I = 1 - \frac{1}{\Psi_I} - \sqrt{\frac{\phi_I}{\Psi_I}} + O(\gamma) \quad \text{and} \quad \rho_{II} = 1 - \frac{1}{\Psi_I} - \sqrt{\frac{\phi_{II}}{\Psi_{II}}} + O(\gamma).
\]

It follows immediately from Equation 32 that \( \lim_{c \to 0^+} \rho_I = \lim_{c \to 0^+} \rho_{II} = 1 \). Using a Taylor series expansion around \( c = 0 \) yields

\[
\rho_I = 1 - \left( \frac{\gamma^2}{\alpha \tau^2} + \frac{\tau^2}{\alpha} \right) + O(\gamma^3) \quad \text{and} \quad \rho_{II} = 1 - (\gamma^2 + \gamma \xi) + O(\gamma^3)
\]
As $c$ is small, we drop $O\left(\gamma^3\right) = O\left(c^2\right)$ from the derivations below.

**Step III: Characterization of $\beta^I_1(\mu_s) = 1$ and $\beta^{II}_1(\mu_h) = 1$ for $c \to 0^+$.** By definition, $\rho_I = \beta^I \xi$. We have that

$$
\beta^I(\mu_s) = 1 \Leftrightarrow \rho_I = \xi \Leftrightarrow \xi^2 = 1 - \left(\frac{\gamma^2}{\alpha^2} + \frac{\gamma \xi}{\alpha^2}\right)
$$

$$
\Leftrightarrow \gamma = \tau \left(-\frac{1}{2} \xi + \frac{1}{2} \sqrt{4\alpha + \xi^2 - 4\alpha \xi^2}\right)
$$

As $\rho = \frac{\Delta}{\mu_1}$, then, we can rewrite the latter inequality as

$$
c V \mu_1 = \frac{\alpha + \rho (1 - \alpha)}{\alpha^2} \left(\frac{1}{2} \sqrt{\rho} + \sqrt{\alpha + \rho \left(\frac{1}{4} - \alpha\right)}\right)^2
$$

$$
= (1 - \rho)^2 - \alpha (1 - \rho)^3 + O\left((1 - \rho)^4\right).
$$

The last step is done by a Taylor series expansion of the right hand side around $\rho = 1$. Thus, we obtained that $\frac{c}{V \mu_1} = (1 - \rho)^2 + O\left((1 - \rho)^3\right)$. Similarly, we have that

$$
\beta^{II}(\mu_h) = 1 \Leftrightarrow \xi^2 = 1 - \gamma^2 + \gamma \xi
$$

$$
\Leftrightarrow \gamma = -\frac{1}{2} \xi + \frac{1}{2} \sqrt{1 - 3\xi^2}
$$

from which, we obtain

$$
c V \mu_1 = \left(-\frac{1}{2} \sqrt{\rho} + \sqrt{1 - \frac{3}{4} \rho}\right)^2
$$

$$
= (1 - \rho)^2 - (1 - \rho)^3 + O\left((1 - \rho)^4\right)
$$

The last step is done by a Taylor series expansion of the right hand side around $\rho = 1$. Thus, we obtained that $\frac{c}{V \mu_1} = (1 - \rho)^2 + O\left((1 - \rho)^3\right)$.

**Proof of Proposition 12.** We can use the results of Lemma 14 to determine the profit rate difference.

When $\beta_1 = 1$, for the honest service strategy case, we obtain $\rho_1 = \rho_0$. As $\alpha = 1$, with honest service we obtain

$$
\frac{\Pi(\mu_h)}{c} = sw_1 - l_1 = v - \left(1 - \rho_1\right) \rho_1.
$$

For the service inducement strategy case, we obtain $\rho_1 \left(\frac{1}{\alpha} - 1\right) = \rho_0 - \rho_1$ and thus

$$
\frac{\Pi(\mu_s)}{c} = sw_1 = v - \rho_0 - \frac{1}{1 - \rho_1} + (\rho_0 - \rho_1)\frac{1 - \rho_1 + \rho_0}{1 - \rho_1 + \rho_0}.
$$
Taking the difference of the two expressions above, we obtain for \( c > 0 \).

\[
\frac{\hat{\Pi}(\mu_h) - \hat{\Pi}(\mu_s)}{c} = \left( \frac{1}{1 - \rho_1} + 1 \right) \rho_1 - \rho_0 \frac{1}{1 - \rho_1 + \rho_0} + \rho_0 - \rho_1
\]

and \( \rho_0 = \frac{1}{\alpha} \rho_1 \), from which we obtain

\[
\frac{\hat{\Pi}(\mu_h) - \hat{\Pi}(\mu_s)}{c} = -1 + \frac{1 + \frac{1}{\alpha} \rho_1 - 2 \rho_1 - 2 \alpha + \alpha \rho_1}{1 + \frac{1}{\alpha} - \frac{1}{\alpha} \mu_1^2} = -1 + \frac{1 + \frac{1}{\alpha} \mu_1 - 2 \mu_1 - 2 \alpha + \alpha \frac{1}{\mu_1}}{1 + \alpha \frac{1}{\mu_1} - 2 \alpha + \alpha \frac{1}{\mu_1}}.
\]

It follows immediately from \( \rho_1 = \frac{\Delta}{\mu_1} \) that the threshold base load is determined by \( \bar{\rho} = \alpha \frac{\xi - 1}{(\alpha - 1) \tau} \).

**Proof of Proposition 13.**

**Overview:** In step I, we determine profit rate expressions with a skimming service strategy and an honest service strategy; \( \hat{\Pi}(\mu_s) \) and \( \hat{\Pi}(\mu_h) \). In step II, we develop the Taylor series expansion for the profit difference for small values of the waiting cost. Finally, in step III, the threshold base load is determined by solving the Taylor series of the profit difference for \( \rho \).

**Step I: Profit rate expressions.** With Equations 27 and 29 of Proposition 11, we obtain that

\[
\hat{\Pi}(\mu_s) - \hat{\Pi}(\mu_h) = \frac{\left( 1 + \rho_1^I \left( \frac{\xi}{\alpha} - 1 \right) \right) V \Lambda - c \frac{\alpha}{\alpha} \left( \frac{1}{1 - \rho_1^I} + \rho_1^I \left( \frac{\xi}{\alpha} - 1 \right) \right) - \left( \frac{1}{\rho_1} V \Lambda - c \left( 1 + \frac{1}{1 - \rho_1^I} \right) \right) \rho_1}{1 - \rho_1^I + \frac{\rho_1^I}{\alpha}}
\]

where \( \rho_0^I = \Delta \rho_0 / \mu_0 \) and \( \rho_1^I = \frac{\Delta}{\mu_1} \) and \( \alpha = \frac{\rho_0}{\mu_0} \), \( \rho_1^I = \frac{\Delta}{\mu_1} \) and \( \rho_1 = \frac{\Delta}{\mu_1} \). \( \rho_1^I \) and \( \rho_1^I \) are determined by Equations 28 and 30.

**Step II: Taylor series expansion for small \( c \).** We can rewrite the profit difference as follows:

\[
\hat{\Pi}(\mu_s) - \hat{\Pi}(\mu_h) = \frac{\left( 1 + \rho_1^I \left( \frac{\xi}{\alpha} - 1 \right) \right) V \Lambda - c \frac{\alpha}{\alpha} \left( \frac{1}{1 - \rho_1^I} + \rho_1^I \left( \frac{\xi}{\alpha} - 1 \right) \right) - \left( \frac{1}{\rho_1} V \Lambda - c \left( 1 + \frac{1}{1 - \rho_1^I} \right) \right) \rho_1}{1 - \rho_1^I + \frac{\rho_1^I}{\alpha}}
\]

\[
= \frac{A}{1 - \rho_1 + \frac{\rho_1}{\alpha}}
\]

We now elaborate terms \( A \) and \( B \). With Equation 33 and with \( \rho_1 = \xi^2 \) we obtain

\[
A = \left( 1 - \left( 1 - \rho_1^I \right) \left( \frac{\xi}{\alpha} - 1 \right) \right) \frac{1}{\alpha \left( 1 - \rho_1^I + \rho_1 \right)} \rho_1 V \mu_1 + O \left( \gamma^3 \right)
\]

\[
= \left( 1 - \frac{2}{\alpha} + \frac{2 \xi}{\alpha} \right) \left( \frac{\xi}{\alpha} - 1 \right) \frac{1}{\gamma^2 + \gamma \xi + \xi^2} \frac{1}{\gamma^2 + \gamma \xi + \xi^2} V \mu_1 + O \left( \gamma^3 \right).
\]

Developing \( A \) with a Taylor series expansion around \( \gamma = 0 \), we obtain

\[
A = \gamma \frac{\alpha \tau - \alpha - \xi^2 + \alpha \xi^2}{\alpha \tau \xi} V \mu_1 + O \left( \gamma^3 \right).
\]

\[
(35)
\]
For $B$, with $c = V\mu_1\gamma^2$, we obtain
\[
B = V\mu_1\gamma^2 \left( \frac{1 + \frac{1}{1 - \rho I}}{1 - \rho I + \alpha \rho_0} - \frac{\rho I (\frac{1}{\alpha} - 1) \rho_0}{1 - \rho I + \rho_0} \right)
\]
\[
= V\mu_1\gamma^2 \left( \frac{\alpha \rho_0}{1 - \rho I + \alpha \rho_0} - \frac{\rho I (\frac{1}{\alpha} - 1) \rho_0}{1 - \rho I + \rho_0} + \frac{1}{1 - \rho I + \rho_0} - \frac{1}{1 - \rho I + \rho_0} \right)
\]
\[
= V\mu_1\gamma^2 \left( 2 - \frac{1}{\alpha} \right) + \gamma^2 \left( \frac{1}{1 - \rho I + \rho_0} - \frac{1}{1 - \rho I + \rho_0} \right). \tag{36}
\]

Now, we elaborate $C$:
\[
C = \gamma^2 \left( \frac{1}{1 - \rho I + \rho_0} \right) - \frac{1}{1 - \rho I + \rho_0} \frac{\alpha}{\rho_0}
\]
\[
= \gamma^2 \left( \frac{1}{(\gamma^2 + \gamma \xi)} \right) + 1 \left( \frac{1}{\alpha \rho_0} \right) + O (\gamma^3).
\]

Developing $C$ with a Taylor series expansion around $\gamma = 0$, we obtain
\[
C = \frac{\gamma}{\xi} (1 - \alpha \tau) + \frac{\gamma^2}{\xi^2} (\alpha - 1) (\xi^2 + 1) + O (\gamma^3)
\]
which can be plugged back into Equation 36:
\[
B = V\mu_1 \left( \gamma^2 \left( 2 - \frac{1}{\alpha} \right) + \frac{\gamma}{\xi} (1 - \alpha \tau) + \frac{\gamma^2}{\xi^2} (\alpha - 1) (\xi^2 + 1) \right) + O (\gamma^3) \tag{37}
\]

Summing $A$ and $B$ obtained in Equations 35 and 37, we obtain with Equation 34 that
\[
\tilde{\Pi} (\mu_s) - \tilde{\Pi} (\mu_h) = \left( \frac{2 \alpha \tau - \alpha - \xi^2 + \alpha \xi^2 - \alpha^2 \tau^2}{\alpha \tau} \frac{\gamma}{\xi} + \frac{\gamma^2}{\xi^2} (\alpha - 1) (\xi^2 + 1) \right) V\mu_1 + O (\gamma^3)
\]

Introducing $\tau^2 = \frac{\alpha + \xi^2 - \alpha \xi^2}{\alpha^2}$,
\[
\tilde{\Pi} (\mu_s) - \tilde{\Pi} (\mu_h) = \left( 2 \left( 1 - \frac{\alpha + (1 - \alpha) \xi^2}{\alpha \tau} \right) \frac{\gamma}{\xi} + \left( 2 - \frac{1}{\alpha} \right) \left( \xi^2 - (1 - \alpha) (\xi^2 + 1) \right) \frac{\gamma^2}{\xi^2} \right) V\mu_1 + O (\gamma^3)
\]
and substituting back $c' = \frac{c}{\mu_1}$, $\gamma^2 = \frac{\gamma}{\mu_1}$, and $\xi^2 = \frac{\xi}{\mu_1}$, we finally obtain that
\[
\tilde{\Pi} (\mu_s) - \tilde{\Pi} (\mu_h) = \left( 2 \left( 1 - \frac{\alpha + (1 - \alpha) \rho}{\sqrt{\alpha + (1 - \alpha) \rho}} \right) \frac{\gamma}{\sqrt{\rho}} + \left( 2 - \frac{1}{\alpha} \right) \frac{\rho - (1 - \alpha) (\rho + 1)}{\rho} \frac{\gamma^2}{\rho} \right) V\mu_1 + O (\epsilon^2)
\]

**Step III: determination of the threshold base load:** Finally, $\tilde{\Pi} (\mu_s) = \tilde{\Pi} (\mu_h) + O (\epsilon^2)$ when $c' = 0$, or, when
\[
\sqrt{c'} = \frac{2 \left( 1 - \frac{\alpha + (1 - \alpha) \rho}{\sqrt{\alpha + (1 - \alpha) \rho}} \right) \sqrt{\rho}}{\left( 2 - \frac{1}{\alpha} \right) \rho - (1 - \alpha) (\rho + 1)}
\]

54
The right hand side can be developed with a Taylor series expansion around $\rho = 1$:

$$
\sqrt{c'} = (1 - \rho) \frac{\alpha - 1}{2 \alpha - 1} + O \left( (1 - \rho)^2 \right) \text{ or } \frac{1}{\alpha - \frac{2}{2 - \alpha}} \sqrt{c'} = (1 - \rho) + O \left( (1 - \rho)^2 \right)
$$