Overselling Expert Services to Waiting Customers

Timothy Chu  
School of Business and Management, Hong Kong University of Science and Technology, ttcchu@ust.hk

Hongtao Zhang  
School of Business and Management, Hong Kong University of Science and Technology, imhzhang@ust.hk

May, 2014

Abstract. Experts possess expertise that may be difficult or unrealistic for clients to acquire. This information asymmetry means that experts could oversell unnecessary services when clients seek their advice. This work considers overselling expert services in a queuing setting. We find that with ample demand, proportional pricing can eliminate overselling. For the case of more than proportional pricing, when customers’ service value is high enough, the expert may want to oversell. We also find that when service times for different services are closer, an expert will exploit the opportunity to oversell more. With limited demand, we find that it is optimal for the expert to have some level of overselling, since the expert will have unused capacity otherwise, and overselling will not drive away customers. This analysis of single queues also provides building blocks for future work on overselling under competition, and the design of mechanisms that could induce truthful services.

(Key words: overselling, waiting time, valuation)

1 Introduction

We seek advice and services from experts — doctors, lawyers, repair technicians, etc. — when we have problems that require specific knowledge to be resolved. The expert possesses expertise that may be difficult or unrealistic for clients to acquire, and this knowledge gap results in information asymmetry between the expert and the client about the extent of the service that is truly needed, even after the service is provided. This is a distinct characteristic of expert service that renders it a “credence good”.

Due to the information asymmetry, experts could oversell unnecessary services to clients. Even though experts across many fields, such as the medical profession, are required to observe ethical codes, which may have references to serving clients honestly, relying solely on the code of ethics may not eliminate the possibility of overselling. Instead, overselling of expert service continues to be a common phenomenon. For example, in 2013, the ABIM Foundation, a medical association in the US, published a list of 90 commonly overused medical procedures such as CT...

This work aims to complement existing literature by considering expert services in a queuing setting. Although the setting can be generally applied to many kinds of expert services, our primary motivating examples will be from the healthcare sector, where patients choose competing healthcare providers. We want to address the following three questions.

i. What is the optimal overselling strategy for experts? When an expert provides unnecessary services to the client, the expert can gain more profit from that particular client, but the service time needed would be longer, which reduces the time spent on other clients. An expert who always serves clients truthfully can speed up the service, attracting more clients, but he gives up the opportunity to gain more from each client; an expert who always oversells can bring in more profits per client, although potentially losing some clients to competitors. The optimal overselling strategy for an expert could be anywhere between these two scenarios. It would depend on the extra time needed for and the extra profit gained from overselling, as well as the proportion of clients who are susceptible to overselling. We want to characterize, under different conditions, how the experts would choose the level of overselling. We also want to find conditions under which the experts will choose to be completely truthful. This is an important question from the society’s point of view, since overselling provides no economic value to clients, causing an inefficient use of resources.

ii. What proportion of the potential client base could be served? Clients face the question of whether to seek for an expert service (joining a queue). Their decisions would be based on how much they value the service provided by the experts, and their expected waiting time (which depends on the overselling strategy of the expert). When overselling occurs, the expected waiting time will be longer and there will be fewer clients willing to queue up. We will examine the proportion of client base that could have been served but are not, i.e., those who decide to drop out because of overselling. In other words, what is the harm of overselling?

iii. What mechanisms could induce the experts to provide truthful services? Overselling could indeed be a very serious issue for the social planner. For example, in Hong Kong, with an ageing population, the government forecasts an increase in healthcare expenditure from 5% of GDP in 2004 to 9% of GDP by 2030 ("Healthcare Reform in Hong Kong", Food and Health Bureau, March 2013). Inefficient use of medical resources from overprovision or overselling could put even more stress on the healthcare system. Consider the point of view of a government or social planner who cares about social welfare maximization and intends to induce the experts not to oversell. We mentioned the code of ethics that helps to mitigate overselling, but in an extreme scenario when experts ignore the ethical code, is it still possible to attain complete honesty of service? Methods may include designing a compensation structure or designing a system that
allocates clients based on performance (e.g., average waiting time).

This paper draws upon two streams of literature.

**Economics of Credence Goods.** Darby and Karni (1973) coined the term credence goods, developed simple models, and identified the existence of overselling. Glazer and Hassin (1983) investigated cheating by taxi drivers. They found that in the absence of non-linear pricing schemes, drivers would make the ride longer than necessary. Also, higher service capacity increases the level of cheating. Emons (1997) considered competition of capacity constrained experts in a deterministic setting and found that competition induces honest diagnosis. Emons (2001) again has a deterministic setting, but with a monopoly expert. The result is that the expert can signal honest diagnosis by setting service capacity exactly equal to market demand, or by charging a fixed admission price, and that it is optimal to do so. Fong (2005) also has a similar setting, with the expert first announcing prices, then observes customer’s problem and recommends treatment; his model results in honesty.

Our model is richer than those above, as we consider a stochastic queueing setting where customers play a role in their queue joining strategies. Also, we model a situation where the expert does not set prices and we focus on the link between overselling, expert’s profit and customers’ welfare.

Dulleck and Kerschbamer (2006) propose a unifying model of credence goods, which showed that customer homogeneity (an assumption in Emons’ studies) is one of the necessary conditions to eliminate overselling. The idea is that with homogeneous customers, a single price leaves customers indifferent between purchasing service or not. In this paper we model heterogeneous clients, in the sense that they need different levels of services.

**Queueing for Expert Services.** Deviating from the deterministic setting in the economics literature, Debo et. al (2008) investigated how the level of overselling is influenced by workload dynamics. In their model, the server’s workload fluctuates with stochastic inter-arrival times and service times as captured by a changing queue length. Although their model has homogeneous customers, the monopoly expert can take advantage of the heterogeneity from customers arriving at different system states (an empty vs. busy queue), and treat them differently. They find that it is never optimal to oversell to customers who enter a busy queue, but it is sometimes optimal to oversell to customers who enter an empty queue. If waiting cost is high, however, then it is never optimal to oversell.

Whereas customers in Debo et. al (2008) are heterogeneous in arriving states, Paç and Veeraraghavan (2010) explicitly model a more realistic setting based on medical services, where heterogeneous customers have different needs but the same symptoms, and customers may decide to join a queue for treatment after he gets a diagnosis from the server. In their work, prices serve two purposes: firstly, a higher price signals the credibility of the expert’s diagnosis; secondly,
prices control congestion. The economics literature showed that a fixed admission fee induces honest service; in Paç and Veeraraghavan (2010), an expert may achieve honest price discrimination, for example, when there is large potential demand, making it less costly to signal honest service.

Whereas Debo et. al (2008) concerns how the expert may oversell depending on workload dynamics in an observable queue, here the queue is unobservable. Paç and Veeraraghavan (2010) are concerned with a price setting monopoly expert. In this paper, we model a situation where the expert is not the price setter but attempts to gain by overselling his service to clients who do not need the extra service.

Anand et. al (2011) considered customer-intensive services, and identified a tradeoff between service speed and service quality. Customer-intensive services, such as healthcare and beauty care, are closely related to expert services. The difference is that in customer-intensive services, there is no information asymmetry between customer and server: when the server spends more time with a customer, both parties know that the value of the service to the customer will increase; but in expert services, overselling cannot provide extra value to customers.

The rest of this paper is organised as follows: Sections 2 and 3 introduce the model and formulate the problem, respectively. Section 4 is the analysis of a queue with ample demand. Section 5 deals specifically with more than proportional pricing structure under ample demand. Section 6 is the analysis for limited demand; we conclude in Section 7.

2 The Model

Customers in need of expert service arrive in the market according to a Poisson process at an exogenous arrival rate of $\Lambda$. Each customer has a problem that requires expert treatment. There are two types of problems, namely the high level/major problem, $H$, and the low level/minor problem, $L$. A customer’s problem is $H$ with probability $\theta$ and $L$ with probability $1 - \theta$, where $\theta \in (0, 1)$ is common knowledge. Since the customer (he) does not have expert knowledge, and the symptoms of $H$ and $L$ problems are similar, he does not know which problem he actually has.

Meanwhile, an expert can instantaneously and correctly diagnose which problem the customer has. There is a high level ($H$) procedure that can resolve both $H$ and $L$ problems, and a low level ($L$) procedure that can only resolve the $L$ problem. Therefore when facing an $H$ customer, the expert must provide the $H$ procedure; but for an $L$ customer, the expert can provide either
the $H$ or $L$ procedure.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Issue that could be resolved</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>$L$</td>
</tr>
<tr>
<td>$L$</td>
<td>$L$ or $H$</td>
</tr>
</tbody>
</table>

Overselling occurs when the expert provides $H$ procedure to an $L$ customer. We assume that experts’ diagnosis is not verifiable, so experts cannot be “caught” overselling. What is verifiable is whether the customer’s problem has been resolved. The expert can choose the probability of overselling to an $L$ customer, $\alpha$, and we shall call this his overselling strategy. Furthermore, the proportion of $L$ customers, $1 - \theta$, captures the level of opportunity for the expert to oversell.

With overselling strategy $\alpha$, we define the probability of the expert performing an $H$ procedure as $\phi = \theta + \alpha(1 - \theta)$, and the expert performs an $L$ procedure with probability $1 - \phi = (1 - \alpha)(1 - \theta)$. Note that $\alpha$ and $\phi$ has a one to one relationship. We call $\phi$ the level of overselling.

For simplicity, we assume that the procedures are standardized, so that the service times are fixed at $t_H$ and $t_L$ for $H$ and $L$ procedures respectively, where $t_H > t_L$. The expected service time for a client is hence $E(T) = \phi t_H + (1 - \phi) t_L$. We may relax this assumption in the future, so that service times can have general distributions.

Customers are charged $p_H$ and $p_L$ for receiving an $H$ and $L$ procedure, respectively. For now, we assume that the expert is a price taker who cannot alter these prices. This is reasonable, for example, in the case where a doctor is employed by a hospital or a healthcare maintenance organization, and the employer decides the pricing; or that there is a standardized price in public hospitals for procedures such as the CT scan. The expert’s cost of providing services (apart from the time spent) is a constant which can be normalized to 0. Hence the expected profit from a customer is $P = \phi p_H + (1 - \phi) p_L$.

Let $v_H$ ($v_L$) be the value that an $H$ ($L$) customer gains from having his problem resolved, and let $C$ be the per unit waiting cost for all customers. Define $V = \theta v_H + (1 - \theta) v_L$ as the client’s expected value for a service. Although the customers do not know their true type, their valuation of a problem getting resolved will not be altered by which procedure they have received. This captures the distinct characteristic of expert services. For example, a patient with a minor head injury ($L$ problem) can be treated with the evaluation of a CT scan ($H$ procedure) or just by clinical observation ($L$ procedure). The value of recovering from the injury will still be $v_L$, no matter which procedure is used.

The following is a summary of notation:

- $\Lambda$ = Potential arrival rate
- $\lambda$ = Actual arrival rate
- $\theta$ = Proportion of $H$ customers
- $v_i$ = Customer’s valuation of service $i = H$ or $L$
3 Problem Formulation

Customers are willing to join the queue as long as the expected net benefit of joining is non-negative, i.e.,

\[ V - P \geq CW, \]

where \( W \) is the expected waiting time in queue. We assume that time in service is not perceived as costly by the customer. Since we have an \( M/G/1 \) queue, by the Pollaczek-Khintchine formula,

\[ W = \frac{\lambda E(T^2)}{2(1 - \lambda E(T))}. \]

Here \( \lambda \) is the customers’ arrival rate, and \( E(T^2) = \phi t_H^2 + (1 - \phi)t_L^2 \) is the second moment of service time. In equilibrium, the expert will set an overselling strategy in order to maximize her profit \( \Pi \), subject to having customers joining the queue,

\[ \max_{\alpha} \Pi = \lambda P \quad \text{s.t.} \quad V - P \geq \frac{C \lambda E(T^2)}{2(1 - \lambda E(T))}. \]

Customers decide whether to join or not with a probability \( \delta \) in order to receive non-negative expected benefit. This results in the arrival rate of \( \lambda = \delta \Lambda \leq \Lambda \). To make that decision, customers will need to infer \( \alpha \) from the expected waiting time \( W \). The information of \( W \) may be obtained through repeated trials, or in the case of a public service, it may be publicly announced.

4 Ample Demand

We first consider the case of ample demand, which is when potential arrival rate is no less than the equilibrium arrival rate from an infinite customer population. A more rigorous definition for ample demand is given in Section 6.
4.1 Properties of $W$, $\lambda$, and $\Pi$

The properties of the expected waiting time, $W$, and the arrival rate $\lambda$, are presented in the following lemmas. Using these results, we can analyse how the expert’s profit change with the level of overselling, $\phi$.

**Lemma 1** The expected waiting time $W(\phi, \lambda)$ is (i) convex and increasing in the level of overselling, $\phi$; (ii) convex and increasing in the arrival rate, $\lambda$. (iii) $\frac{\partial^2 W}{\partial \lambda \partial \phi} > 0$.

**Lemma 2** The arrival rate $\lambda(\phi)$ is decreasing and concave in the level of overselling, $\phi$.

In other words, with more overselling, fewer customers are willing to join, and the rate at which the expert loses customers becomes greater.

**Proposition 1** The expert’s profit $\Pi(\phi)$ is a concave function.

As $\phi$ increases, the congestion effect, $\lambda'(\phi)P(\phi)$, becomes more negative since $\lambda''(\phi)P(\phi) + \lambda'(\phi)P'(\phi) < 0$, and the revenue effect, $\lambda(\phi)P'(\phi)$, becomes less positive since $\lambda'(\phi)P'(\phi) < 0$.

4.2 Profit Maximization

From Proposition 1 we know that $\Pi(\phi)$ is strictly concave, hence $\Pi'(\phi) = 0$ has a unique solution, if it has a solution at all. Define $\hat{\phi}$ to be the maximizer of $\Pi(\phi)$. We call $\hat{\phi}$ the overselling threshold.

**Lemma 3** $\frac{E(T^2)}{E(T)}$ increases in $\phi$.

Here we define fee structures:

**Definition 1** The expert charges (i) proportional fees if $\frac{p_L}{p_H} = \frac{t_L}{t_H}$; (ii) less than proportional fees if $\frac{p_L}{p_H} > \frac{t_L}{t_H}$; (iii) more than proportional fees if $\frac{p_L}{p_H} < \frac{t_L}{t_H}$.

**Lemma 4** $\frac{P}{E(T)}$ is (i) constant in $\phi$ for proportional fees; (ii) decreasing in $\phi$ for less than proportional fees; and (iii) increasing in $\phi$ for more than proportional fees.

**Proposition 2** For proportional and less than proportional fees, $\hat{\phi} = 0$ and the expert maximizes profit by setting $\phi = 0$, i.e., the expert does not oversell.

This Proposition indicates that when the expert charges proportional or less than proportional fees, there is no incentive to oversell, since the congestion effect dominates the revenue effect. It is unclear though, whether or not an expert will have an incentive to oversell in the case of more than proportional fees. Since in that case, both $\frac{P}{E(T)}$ and $\frac{CE(T^2)}{V-P}$ are increasing in $\phi$. We devote the next section for the case of more than proportional fees.
5 More Than Proportional Fees

As defined earlier, 
\[ \lambda(\phi) = \frac{2}{f(\phi)}, \]
where 
\[ f(\phi) = \frac{CE(T^2)}{V - P} + 2E(T), \]
which increases in \( \phi \). We can express \( \Pi'(\phi) \) as
\[ \Pi'(\phi) = P(\phi)\lambda'(\phi) + P'(\phi)\lambda(\phi) = -\frac{2}{(f(\phi))^2}g(\phi), \]
where \( g(\phi) = P(\phi)f'(\phi) - P'(\phi)f(\phi) \). Note that \( g(\phi) \) is strictly increasing since \( \Pi'(\phi) \) is strictly decreasing by the fact that \( \Pi(\phi) \) is strictly concave.

Lemma 5 (i) If \( g(0) > 0 \), then \( \hat{\phi} = 0 \), and \( \alpha^* = 0 \).
(ii) If \( g(0) < 0 \) and \( g(1) > 0 \), then \( g(\hat{\phi}) = 0 \). Further, if \( \theta < \hat{\phi} \), then it is optimal to oversell at \( \phi^* = \hat{\phi} \), i.e., \( \alpha^* = \frac{\hat{\phi} - \theta}{1 - \theta} \). If \( \theta \geq \hat{\phi} \), then it is optimal not to oversell, \( \phi^* = \theta \), i.e., \( \alpha^* = 0 \).
(iii) If \( g(1) < 0 \), then \( \hat{\phi} = 1 \), i.e., \( \alpha^* = 1 \), it is optimal to always oversell.

We next consider cases when some level of overselling is optimal.

**Proposition 3** (i) There exists a threshold \( V_0 \) of customers’ service value \( V \), above which the overselling threshold \( \hat{\phi} > 0 \). (ii) There exists a threshold \( V_1 \) of customers’ service value \( V \), under which the overselling threshold \( \hat{\phi} < 1 \). (iii) \( V_0 < V_1 \).

To summarise, when \( V < V_0 \) (\( g(0) \) and \( g(1) > 0 \)), it is optimal not to oversell; when \( V_0 < V < V_1 \) (\( g(0) < 0 \) and \( g(1) > 0 \)), then \( 0 < \hat{\phi} < 1 \); when \( V > V_1 \) (\( g(0) < g(1) < 0 \)), then it is optimal to always oversell.

**Technical note:** in (ii) we assumed \( V > p_H \), so that \( \hat{\phi} \) can take any value between \( 0 < \hat{\phi} < 1 \). We can relax this assumption. If \( p_L < V < p_H \), there is a \( \tilde{\phi} \) such that \( P(\tilde{\phi}) = p_L + (p_H - p_L)\tilde{\phi} = V \), i.e., \( \tilde{\phi} = \frac{V - p_L}{p_H - p_L} \). Our restriction for \( \tilde{\phi} \) is then \( 0 < \tilde{\phi} < 1 \).

5.1 Effect of Service Value on the Overselling Threshold

From the previous proposition we know that when \( V \in (V_0, V_1) \), the overselling threshold \( \hat{\phi} \in (0,1) \). We aim to find out whether \( \hat{\phi} \) increases monotonically with \( V \), i.e., when customers’ valuation increases, does it always incentivize the expert to oversell more?

**Proposition 4** \( \hat{\phi} \) is increasing in \( V \in (V_0, V_1) \).
5.2 Effect of Service Time on the Overselling Threshold

Following Proposition 2, we know that when the expert charges more than proportional fees, it is possible that $\hat{\phi} > \theta$. Here we focus on the situation where $g(\hat{\phi}) = 0$ for $0 < \hat{\phi} < 1$, that is, $g(0) < 0$ and $g(1) > 0$, namely, $\Pi'(0) > 0$ and $\Pi'(1) < 0$.

**Proposition 5** As the service time of the two services gets closer, the more likely the expert will oversell. That is, $\hat{\phi}$ increases when either $t_L \to t_H$ or $t_H \to t_L$.

The result of this proposition is intuitive for $t_H$ approaching $t_L$. When $t_H$ decreases, the congestion effect will decrease, and the cost of overselling is reduced, i.e., the customer will not mind as much. But when $t_L$ approaches $t_H$, although the extra cost of being oversold decreases (from the customer’s perspective), there will be more congestion. This proposition shows that the congestion effect is always less significant than the increased opportunity to oversell.

6 Limited Demand

So far we have analyzed the case with ample demand, which is when $\Lambda \geq \lambda(\max(\theta, \hat{\phi}))$. In this case, the optimal level of overselling is $\phi^* = \max(\theta, \hat{\phi})$, the equilibrium arrival rate is $\lambda(\phi^*)$, and $\Pi(\phi^*)$ is the best the expert can achieve, i.e., an increase in $\Lambda$ beyond $\lambda(\max(\theta, \hat{\phi}))$ has no effect on the profit. There are two cases of limited demand to be considered:

**Definition 2** (i) Very limited demand is when $\Lambda < \lambda(1)$.

(ii) Moderately limited demand is when $\lambda(1) \leq \Lambda < \lambda(\max(\theta, \hat{\phi}))$.

With very limited demand, equilibrium arrival rate is $\Lambda$, and the optimal level of overselling is $\phi^* = 1$. The expert can always oversell without driving away customers. The customers retain some surplus in this case. The expert’s profit is $p_H\Lambda$. Customers enjoy a positive surplus,

$$V - P - CW = V - p_H - \frac{CM^2_H}{2(1 - \Lambda t_H)} > 0.$$  

With moderately limited demand, the equilibrium arrival rate is $\Lambda$, and the optimal level of overselling is $\phi^* = \lambda^{-1}(\Lambda) \in (\theta, 1]$, i.e., set to make

$$V - P - CW = V - P - \frac{C\Lambda E(T^2)}{2(1 - \Lambda E(T))} = 0.$$  

The expert can still extract all consumer surplus like with ample demand, and makes a profit of $P(\phi^*)\Lambda$. 

9
7 Conclusion

In this work we have mainly focused on the first question that is set out in the Introductory section: the optimal overselling strategy of experts in a queueing setting. We find that with ample demand and proportional or less than proportional pricing, it is optimal never to oversell. This is because the revenue effect of overselling is always dominated by the congestion effect. With ample demand and more than proportional pricing, it is possible that some level of overselling is optimal. The condition is that customers have to value the service high enough. In which case the revenue effect is high (more than proportional pricing), and the congestion effect is dampened. We also find that the expert will oversell more when the customers’ service valuation is higher or when the service times for the two level of service are closer. Under very limited demand, it is optimal to always oversell. For moderately limited demand, it is optimal to oversell at a certain probability such that the expert extracts all surplus.

Acknowledgments. The authors gratefully acknowledge the support of The Hong Kong Research Grant Council General Research Fund [Grant #644009].

References