Competition, complementarity and service level guarantee in Web services

Ali Habibi Badrabadi¹ and Mohammad Jafar Tarokh²

Network and processing overhead associated with web services is a significant challenge to its performance. As a result, web service providers often announce a service level agreement. This ensures that consumers, who pay for the service, can get the service at a given quality level. In this paper, we study the competition between two providers offering functionally the same web services, where there is a monopoly service provider who offers a service that is complementary to their services. Each provider needs to decide a service level (L or H) he/she would offer and a corresponding price for the selected service level to meet the QoS guarantee. We combine modeling constructs from game theory and queuing theory to propose a model that can provide useful insights to service providers about pricing and general competitive strategies.

Keywords: Web services, WSLA, Pricing, Game theory, Queuing theory, Competition, Complementarity.

1. Introduction

In recent years, web services have gained popularity as a useful and efficient technology for developing and integrating web applications, [16]. Unlike traditional software systems, web services are modular software components that are delivered to the user over a network (such as Internet) and executed on a remote system hosting the requested services. Web services encapsulate certain business functionality, and are usually self-contained, loosely coupled, and programmatically accessible using standard Internet protocols that permit different types of systems to share information without human intervention (Pilioura et al., 2007). This can also form the basis of inter-organizational collaboration and distributed process management (for more information, [4, 12, 17].

Web services provide an inexpensive and rapid solution for system development and integration. Typically, service providers create and publish components with specific functionalities. A service consumer who needs certain functionality can resort to various web service discovery mechanisms to locate and invoke the service using standard protocols by paying a fee, [2, 15]. One characteristic of these services is that the integration needed by the user is no longer prohibitively expensive or time consuming. Therefore, complementarity is an important characteristic of web services, [13]. Independent services can be composed in processes to provide even greater value than the sum of component services [3].

Despite its great promise, there are serious concerns about the performance of a web service because of the network bandwidth and processing overhead associated with transferring the large and complex XML-based messages over the network. In some instances, a service provider may not be able to handle the throughput, resulting in serious performance degradation. As a result, launching a service is often associated with announcing a web service-level agreement (WSLA) along with the capability and interfaces of the service. The service capability and interfaces define the conceptual purpose as well as the input/output of the web service. On the other hand, WSLA defines agreed performance metrics and ways to evaluate and measure them, [6, 9, 15]. The role of WSLA is shown in Figure 1.

A service-level agreement could include several quality metrics, such as response time or latency, availability, accessibility, reliability, and versioning, [1, 6]. Among the various quality dimensions, response time is often considered to be the most critical dimension and the most
difficult one to manage. Hence, it is the focus of our work here. Response time refers to how long it takes that a web service responds to the request of a user and is typically measured as an average time over a specified time horizon.

Irrespective of the robustness or the functionalities of a web service, in order to maintain a WSLA, specially with respect to the response time, it is necessary for the web service provider to design and implement a proper pricing scheme. A pricing scheme works as an efficient access control mechanism by providing sufficient incentives to users to choose appropriate service levels.

Although there is a rich body of existing research on pricing products and services, these traditional pricing schemes do not usually work well for web services: First, the marginal cost of providing the service to an additional user is negligible, thereby reducing the traditional price to zero. Second, a very important aspect here is the social cost of congestion; traditional pricing models do not capture this negative externality. On the other hand, non-pricing approaches to access control for reducing the congestion cost are either flawed or, more generally, have undesirable side effects (MacKie-Mason and Varian, 1994 and Zhang et al., 2009).

The main contribution our work is the development of a framework for competitive pricing of web services. We combine modeling constructs from game theory and queuing theory to propose a rigorous model that can provide useful insights to service providers about pricing, and general competitive strategies. More specifically, we study duopoly competition between two providers offering web services with the same functionality, while there is a monopoly service provider who offers a web service that is complementary to their services. Facing a continuum of users who value the benefits from the service against its price (plus the delay cost), a provider needs to decide a service level (L or H) she would offer, and a corresponding price for the selected service level to meet the QoS guarantee (in terms of the average response time of the service).

The remainder of the paper is organized as follows. First, in Section 2, we describe the basic modeling framework, and then in Section 3 we analyze the equilibrium pricing strategy when two competitors choose service levels and prices simultaneously, while the monopoly as a leader had chosen service level H. Section 4 considers equilibrium strategies using results of the model simulation. Finally, section 5 concludes the paper and offers future research directions.

2. The model

Assume that there are three service providers: (1) Two identical providers, indexed 1 and 2, offering web services (or software components) with the same functionalities, S1, both of them having fixed processing capacity, (2) a monopoly service provider indexed 3, with an unlimited processing capacity offering S2 that is complementary to S1. The target market is divided into three segments: customers who are interested in S1 only (Segment 1), customers who need both S1 and S2 (Segment 2) and customers who want S2 only (Segment 3). Figure 2 shows this
market structure. The services can be delivered in one of two discrete levels, \(L \) and \(H \), each level is characterized by an expected total response time of the service. Let \(d_L \) and \(d_H \) be respectively the guaranteed response time for the service offered with level \(L \) and \(H \), where \(d_L > d_H \). We assume that \(d_L \) and \(d_H \) are exogenous and fixed. Provider \(i, (i = 1, 2, 3) \), can choose either \(L \) or \(H \) service level and charges a corresponding price, \( P_i(j) (j = L \ or \ H) \), for it. Each provider QoS guarantees that the actual expected response time for his/her service will be at most \(d_j \). Since the marginal cost of providing a web service is negligible, we skip it and assume costs of developing web services are sunk.

![Figure 2. Market segments and the total arrival of their users looking for web services](image)

The users are characterized by two parameters, \(v \) and \(h \). More specifically, \(v \) denotes the value that user assigns to the immediate provision of each single service while \(h \) denotes the disutility (cost) incurred by the user when the provision of service is delayed by one time unit. For simplicity we assume the value of \(v \) that each user assigns to \(S1 \) and \(S2 \) are equal. The disutility is assumed to be a linear function of \(h \); i.e. if the delay-sensitivity is \(h \), then the disutility is \(h\gamma \) per time unit of delay, where \(\gamma \) is a constant. We assume that both \(v \) and \(h \) are uniformly distributed and are normalized in interval \([0, 1] \). For each user of segment 2 who is interested to compose \(S1 \) and \(S2 \), there is a parameter \(v_2 \). \(v_2 \) denotes the value that user assigns to the immediate provision of the composite web service \((v_2 \geq v)\).

In order to simplify our analysis, we assume that \(v_2 \) and \(v \) are linearly related: \(v_2 = \alpha v \), where \(\alpha \geq 1 \). Furthermore, we assume that \(v \) and \(h \) are linearly related: \(h = \delta v + \eta \). This assumption is reasonable, since a high value of \(v \) usually means that the user is also more delay-sensitive in obtaining the service. We should consider \(\eta = 0 \). This is because users with zero valuation of the service \((v = 0) \) would not be sensitive to delay at all \((h = 0) \). Given this, without any loss of generality, we can set \(\delta = 1 \) (because of our assumption about presence of \(\gamma \) as a constant); i.e., \(v = h \).

At the time of requesting a service, the utility of a user, with \(v \) and \(h \), who chooses to obtain \(S1 \) or \(S2 \) from provider \(i \) is equal to: \(v - p_i(j) + h\delta d_j \). Assume that in each segment the arrival of users looking for web services follows a Poisson process with a base rate of \(\lambda_0 \) (in order to simplify our analysis, as seen in Figure 2, we assume that the total arrival rate of users in all of the segments are equal); i.e., \(\lambda_0 \) is the total arrival rate of users when the response time of the service is zero and the service is free. The effective arrival of users accessing service from provider \(i \) also follows a Poisson process, but with a rate \(\lambda_i \) proportional to \(\lambda_0 \) , \([11]\). The service time of each request follows an arbitrary distribution with a mean service time of \(b \) second. In other words, the processing rate of each provider is \(\frac{1}{b} \). The above system can be modeled as an M/G/1 processor-sharing queue, and the expected total response time \(W_i^j \) from provider \(i \) is given by Kleinrock \([5]\) and Zhang et al.\([15]\):

\[
W_i^j = \frac{b}{1 - \rho \frac{\lambda_i}{\lambda_0}}, \quad (1)
\]

where \(\rho \equiv \lambda_0 b > 0 \) represents the normalized total traffic intensity. \(W_i^j \) should not be greater than announced response time of the service, \(d_j (j = L \ or \ H) \).
As mentioned above, each service provider can choose offering his/her service with level L or level H. Therefore, there are eight cases for these three providers. Note that the order of service provider 1 (SP1) and service provider 2 (SP2) is not important. So, in cases which these two providers choose differentiated service levels, without loss of generality, we can assume that SP1 chooses H while SP2 chooses L. Hence, we have six different cases, in three of them the monopoly (SP3) chooses level H and in the others chooses level L. Table 1 shows these six cases.

<table>
<thead>
<tr>
<th>Case</th>
<th>Chosen Service Level</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H H H</td>
<td>HHH</td>
</tr>
<tr>
<td>2</td>
<td>L L H</td>
<td>LLH</td>
</tr>
<tr>
<td>3</td>
<td>H L H</td>
<td>HLH</td>
</tr>
<tr>
<td>4</td>
<td>H H L</td>
<td>HHL</td>
</tr>
<tr>
<td>5</td>
<td>L L L</td>
<td>LLL</td>
</tr>
<tr>
<td>6</td>
<td>H L L</td>
<td>HLL</td>
</tr>
</tbody>
</table>

In the next section we assume that provider 3 chooses level H, and therefore cases 1, 2 and 3 are considered and optimal prices for each provider in these cases are analyzed.

3. Optimal pricing

In our analysis, we assume that service time and total response time of the composite web service follows the high watermark rule. This means that they are equal to maximum of service times and total response times of S1 and S2.

In this section, we consider the situation where provider 3 had offered S2 with level H and price $P_3^{(H)}$ and service providers 1 and 2 choose their service levels and price decisions simultaneously and non-cooperatively. In other words, when making his/her own decisions, a provider cannot observe the choices made by the other. It can be modeled as a Stackelberg game in which one leader (service provider 3) moves first, and decides upon her service level and price and the other providers (followers), observing the choice of the leader, then chooses her/his service level and the corresponding price. The leader anticipates the reaction of the followers, and chooses a price, $P_3^{(H)}$, to maximize her profit, from which we can determine the optimal prices. Note that according to our assumption they are rational and have same information about the market and their preferences.

3.1. Case 1: HHH

When both the competitors offer their services with same service level, consumers of segments 1 and 2 buy from the provider who charges the lowest price. If they both charge the same price (as well as same service level), the probability of demand for each one will be equal to 0.5, and therefore each provider would face a demand equal to the half of the market demand at that price. Because of the capacity constraint and the service-level agreement, however, a provider may not be able to reduce his/her price to the marginal cost. We now investigate the range of prices a service provider is allowed to charge.

Let $P_{jH}$ be the price charged by only one provider offering S1 (when the other competitor is out of the market) such that the response-time constraint due to the service level agreement is
binding. Since \( v \) is uniformly distributed in \([0, 1]\), in segment 1 only a portion of users with \( v - \frac{P_H}{1 - \gamma d_H} > 0 \) or \( v > \frac{P_H}{1 - \gamma d_H} \) and in segment 2 a portion of users with,

\[
v_2 - \frac{P_H}{1 - \gamma d_H} > 0 \quad \text{or} \quad v > \frac{P_H + P_3^{(H)}}{\alpha(1 - \gamma d_H)}
\]

will choose service from the provider (note that \( v_2 = \alpha v \)). In other words, using probability Zhang et al. [15], the effective arrival rate for providers 1 and 2 is:

\[
\lambda_H = \lambda_0 \left( 1 - \frac{P_H}{1 - \gamma d_H} \right) + \lambda_0 \left( 1 - \frac{P_H + P_3^{(H)}}{\alpha(1 - \gamma d_H)} \right).
\]

Hence, \( \underline{P_H} \) solves the following equation:

\[
\frac{b}{1 - \rho(2 - \frac{(\alpha + 1)\underline{P_H} + P_3^{(H)}}{\alpha(1 - \gamma d_H)})} = d_H.
\]

Furthermore, let \( P_H \) be the common price charged by each competitor when they split the market equally and keep the response-time constraint binding. In this situation, the effective arrival rate for each provider can be obtained as \( \lambda_i^{(H)} = \frac{1}{2} \lambda_0 \left( 2 - \frac{(\alpha + 1)P_H + P_3^{(H)}}{\alpha(1 - \gamma d_H)} \right), \ (i = 1, 2) \).

Hence, \( \underline{P_H} \) solves the following equation:

\[
\frac{b}{1 - \frac{1}{2} \rho(2 - \frac{(\alpha + 1)\underline{P_H} + P_3^{(H)}}{\alpha(1 - \gamma d_H)})} = d_H.
\]

Solving the last two equations, we get:

\[
\underline{P}_H = \frac{\alpha(1 - \gamma d_H)}{\alpha + 1} \left( 2 - \frac{u_H}{\rho} \right) - \frac{P_3^{(H)}}{\alpha + 1} \quad \text{and} \quad \underline{P}_H = \frac{2\alpha(1 - \gamma d_H)}{\alpha + 1} \left( 1 - \frac{u_H}{\rho} \right) - \frac{P_3^{(H)}}{\alpha + 1},
\]

where \( u_H = 1 - \frac{b}{d_H} \). It can be easily verified that both \( \underline{P}_H \) and \( \underline{P}_H \), as defined above, can be negative if the total traffic intensity, \( \rho \), is very low. This is because, when \( \rho \) is small, the delay constraint could become slack, and artificially forcing the delay constraint to be binding would lead to negative prices. Of course, prices cannot be negative in reality, so we re-define:

\[
\underline{P}_H = \max \{0, \frac{\alpha(1 - \gamma d_H)}{\alpha + 1} \left( 2 - \frac{u_H}{\rho} \right) - \frac{P_3^{(H)}}{\alpha + 1} \} \quad \text{and} \quad \underline{P}_H = \max \{0, \frac{2\alpha(1 - \gamma d_H)}{\alpha + 1} \left( 1 - \frac{u_H}{\rho} \right) - \frac{P_3^{(H)}}{\alpha + 1} \}.
\]

Lemma 1: When two competitors choose the same service level, (i) the equilibrium price \( P_i^{(j)} \) must satisfy \( P_i^{(j)} \leq P_i^{(j)} \leq \underline{P}_j \), and (ii) every symmetric price choice, \( P_1^{(j)} = P_2^{(j)} = P_j \in [\underline{P}_j, \underline{P}_j] \), is an equilibrium.

Proof: (i) If provider 1 charges a price \( P_1^{(j)} > P_j \), provider 2 can simply set \( P_2^{(j)} = \underline{P}_j \) and get the entire market while meeting the service-level agreement. Hence, provider 1 would be better off by reducing his/her price to \( \underline{P}_j \). On the other hand, provider 1 would not charge a price \( P_1^{(j)} < P_j \). This is because if he/she does, the provider 2 would charge a price \( P_2^{(j)} > P_1^{(j)} \) in order to commit to his/her service level guarantee. All the consumers would then choose service from provider 1, resulting in violation the service level guarantee. (ii) Assume that provider 1
offers a price $P^{(j)}_1 \in [\bar{P}_j, P_j]$. If provider 2 chooses $P^{(j)}_2 < P^{(j)}_1$, all consumers would prefer the service from 2, leading to the violation of provider 2's service level guarantee. Provider 2 also would not choose $P^{(j)}_2 > P^{(j)}_1$ either, because in that case all consumers would choose the service from 1, resulting in zero profit for 2. Hence, the only choice for provider 2 is to charge $P^{(j)}_1 = P^{(j)}_2$. Therefore, any price choice $P^{(j)}_1 = P^{(j)}_2 = P_j \in [\bar{P}_j, P_j]$ is a Nash equilibrium.

From all of the possible symmetric Nash equilibria, we consider only the Pareto-dominant one. Let $P_H$ be the common price charged by both the providers if they both choose service level H. Then, the effective demand to each provider is $\lambda_H = \frac{1}{2} \frac{\lambda_0 (2 - (\alpha + 1)P_H + P_3^{(H)})}{\alpha (1 - \gamma d_H)}$.

Hence, the optimization problem of each provider can be written as:

$$\max_{P_H} \pi^{(H)}_i = \frac{1}{2} \lambda_0 P_H (2 - \frac{(\alpha + 1)P_H + P_3^{(H)}}{\alpha (1 - \gamma d_H)})$$

s.t. $1 - \frac{1}{2} \frac{\rho(2 - \frac{(\alpha + 1)P_H + P_3^{(H)}}{\alpha (1 - \gamma d_H)})}{\alpha (1 - \gamma d_H)} \leq d_H$

$$P_H \leq P_H \leq \bar{P}_H$$

**Lemma 2:** When the monopoly (leader) chooses level H and then both competitors (followers) choose H too, the Pareto optimal price for competitors is given by:

$$P_H = \frac{(1 - \gamma d_H)}{\alpha + 1} \times \begin{cases}
0 & \text{if } \rho < r_1 \\
\frac{\alpha(2\alpha + 3)}{\alpha + 2} - \frac{\alpha^2 + 4\alpha + 1}{\alpha + 2} \frac{u_H}{2\rho} & \text{if } r_1 \leq \rho < r_2 \\
\frac{\alpha(2\alpha + 1)}{2(2\alpha^2 + 4\alpha + 1)} \left( u_H - \frac{2\alpha^2 + 4\alpha + 1}{\alpha + 2} \rho \right) & \text{if } r_2 \leq \rho < r_3 \\
\frac{\alpha(2\alpha + 1)}{\alpha + 2} - \frac{\alpha^2 + 4\alpha + 1}{\alpha + 2} \frac{u_H}{\rho} & \text{Otherwise.}
\end{cases}$$

Also the optimal price for provider 3 (leader) is:

$$P_3^{(H)} = \frac{(1 - \gamma d_H)}{\alpha + 1} \times \begin{cases}
\frac{\alpha}{\alpha + 1} & \text{if } \rho < r_1 \\
\frac{1}{\alpha + 2} \left( \alpha + \frac{u_H}{2\rho} \right) & \text{if } r_1 \leq \rho < r_2 \\
\frac{\alpha(2\alpha + 1)}{2\alpha^2 + 4\alpha + 1} & \text{if } r_2 \leq \rho < r_3 \\
\frac{1}{\alpha + 2} \left( \alpha + \frac{u_H}{\rho} \right) & \text{Otherwise.}
\end{cases}$$
where, \( r_1 = u_H \left( \frac{2\alpha^2 + 4\alpha + 1}{2\alpha(2\alpha + 3)} \right) \), \( r_2 = u_H \left( \frac{(2\alpha^2 + 4\alpha + 1)^2}{2\alpha(\alpha + 1)(2\alpha^2 + 4\alpha + 1) + \alpha(\alpha + 1)(\alpha + 2)} \right) \) and \( r_3 = u_H \left( \frac{2(2\alpha^2 + 4\alpha + 1)^2}{2\alpha(\alpha + 1)(2\alpha^2 + 4\alpha + 1) + \alpha(\alpha + 1)(\alpha + 2)} \right) \).

**Proof:** We can solve (2) to obtain the equilibrium price. The formulation is a simple quadratic optimization problem with one decision variable \( P_H \). So, we have,

\[
P_H = \frac{(1 - \gamma d_H)}{\alpha + 1} \times \begin{cases} 
0 & \text{if } \rho < r_1 \\
2(1 - \frac{u_H}{\rho}) - P_3^{(H)} & \text{if } r_1 \leq \rho < r_2 \\
\frac{P_3^{(H)}}{2} & \text{if } r_2 \leq \rho < r_3 \\
2(1 - \frac{u_H}{\rho}) - P_3^{(H)} & \text{Otherwise.}
\end{cases}
\] (5)

Furthermore, in this case, the effective arrival rate for provider 3 from users of segment 3 and segment 2 can be obtained as \( \lambda_3^{(H)} = \lambda_0 \left( 1 - \frac{P_3^{(H)}}{1 - \gamma d_H} \right) + \lambda_0 \left( 1 - \frac{P_3^{(H)} + P_H}{\alpha(1 - \gamma d_H)} \right) \). Since there is no capacity limitation for provider 3, the optimization problem can be written as:

\[
\max_{P_3^{(H)}} \pi_3^{(H)} = \lambda_0 P_3^{(H)} \left( 2 - \frac{(\alpha + 1)P_3^{(H)} + P_H}{\alpha(1 - \gamma d_H)} \right)
\] (6)

Now, Substituting (5) into (6) and using first order condition of \( \pi_3^{(H)} \) we can obtain \( P_3^{(H)} \) and then substituting it into (5), \( P_H \) is obtained. In order to complete the proof, the thresholds of \( \rho \) can be found by comparing the optimal prices.

### 3.2. Case 2: LLH

Since \( d_L < d_H \), total response time of composite web service for users of segment 2 is equal to \( d_L \). Therefore, if only there is one service provider, his/her effective arrival rate is:

\[
\lambda_L = \lambda_0 \left( 1 - \frac{P_L}{1 - \gamma d_L} \right) + \lambda_0 \left( 1 - \frac{P_L + P_3^{(H)}}{\alpha(1 - \gamma d_L)} \right),
\]

and effective arrival rate of each competitors when they split the market equally and keep the response-time constraint binding is:

\[
\lambda_i^{(L)} = \frac{1}{2} \lambda_0 \left( 2 - \frac{(\alpha + 1)P_L + P_3^{(H)}}{\alpha(1 - \gamma d_L)} \right), \quad (i = 1, 2).
\]

Hence, we have,

\[
\frac{b}{1 - \rho(2 - \frac{(\alpha + 1)P_L + P_3^{(H)}}{\alpha(1 - \gamma d_L)})} = d_L \quad \text{and} \quad \frac{b}{1 - \frac{1}{2} \rho(2 - \frac{(\alpha + 1)P_L + P_3^{(H)}}{\alpha(1 - \gamma d_L)})} = d_L
\]

Solving the above two equations, we get,

\[
\bar{P}_L = \max \{ 0 \}, \quad \frac{\alpha(1 - \gamma d_L)}{\alpha + 1} \left( 2 - \frac{u_L}{\rho} - \frac{P_3^{(H)}}{\alpha + 1} \right) \quad \text{and} \quad \bar{P}_L = \max \{ 0 \}, \quad \frac{2\alpha(1 - \gamma d_L)}{\alpha + 1} \left( 1 - \frac{u_L}{\rho} - \frac{P_3^{(H)}}{\alpha + 1} \right).
\]

Using Lemma 1, in this case the effective demand to each provider is
\[ \lambda_L = \frac{1}{2} \lambda_0 (2 - \frac{(\alpha + 1)P_L + P_3^{(H)}}{\alpha(1 - \gamma d_L)}) \]. Hence, the optimization problem of each provider can be written as:

\[
\max_{P_i} \pi_i^{(L)} = \frac{1}{2} \lambda_0 P_L (2 - \frac{(\alpha + 1)P_L + P_3^{(H)}}{\alpha(1 - \gamma d_L)}) \tag{7}
\]

s.t.

\[
1 - \frac{1}{2} \rho (2 - \frac{(\alpha + 1)P_L + P_3^{(H)}}{\alpha(1 - \gamma d_L)}) \leq d_L
\]

\[
P_L \leq P_L \leq \bar{P}_L.
\]

**Lemma 3** When the monopoly (leader) chooses level \(H\) and then both competitors (followers) choose \(L\), the Pareto optimal price for competitors is given by:

\[
P_L = \frac{(1 - \gamma d_L)}{\alpha + 1} \times \begin{cases} 
0 & \text{if } \rho < r_4 \\
\alpha(2 - \gamma x_i) - (2\alpha + x_i)\frac{u_L}{2\rho} & \text{if } r_4 \leq \rho < r_5 \\
\alpha(1 - \frac{x_i}{2}) & \text{if } r_5 \leq \rho < r_6 \\
\alpha(2 - \gamma x_i) - (2\alpha + x_i)\frac{u_L}{\rho} & \text{otherwise}
\end{cases}
\tag{8}
\]

Also, the optimal price for provider 3 (leader) is:

\[
P_3^{(H)} = (1 - \gamma d_H) \times \begin{cases} 
\frac{1 - \gamma d_H}{\alpha(1 - \gamma d_L) + (1 - \gamma d_H)} & \text{if } \rho < r_4 \\
x_1(\alpha + \frac{u_L}{2\rho}) & \text{if } r_4 \leq \rho < r_5 \\
x_2(\alpha + \frac{u_L}{\rho}) & \text{if } r_5 \leq \rho < r_6 \\
\text{Otherwise},
\end{cases}
\tag{9}
\]

where,

\[
x_1 = \frac{(1 - \gamma d_H)}{(\alpha + 1)(1 - \gamma d_L) + (1 - \gamma d_H)}, \quad x_2 = \frac{\alpha(1 - \gamma d_H)}{2\alpha(\alpha + 1)(1 - \gamma d_L) + (2\alpha + 1)(1 - \gamma d_H)},
\]

\[
r_4 = u_L\left(\frac{2\alpha + x_i}{2\alpha(2 - x_i)}\right), \quad r_5 = u_L\left(\frac{2\alpha + x_i}{\alpha(2 + x_i - 2x_i)}\right) \text{ and } r_6 = u_L\left(\frac{2(2\alpha + x_i)}{\alpha(2 + x_i - 2x_i)}\right).
\]

**Proof:** See proof of Lemma 2. Note that in this case we have,

\[
\lambda_3^{(H)} = \lambda_0 (1 - \frac{P_3^{(H)}}{1 - \gamma d_H}) + \lambda_0 (1 - \frac{P_3^{(H)} + P_L}{\alpha(1 - \gamma d_L)}) = \lambda_0 (2 - \frac{P_3^{(H)}}{1 - \gamma d_H} - \frac{P_3^{(H)} + P_L}{\alpha(1 - \gamma d_L)}).
\]

3.3. Case 3: HLH

Without loss of generality, assume that provider 1 chooses level \(H\) while provider 2 chooses level \(L\). This, of course, means that provider 1 must charge a price \(P_1^{(H)}\) higher than \(P_2^{(L)}\) charged by provider 2.
In order to find the equilibrium price, we need to estimate the expected demand for each provider. For users of segment 1, let \( V \) be the \( v \)-value of the marginal user of segment 1 who is indifferent between two providers. This implies that \( V - P_L^{(i)} - V\gamma d_H = V - P_L^{(2)} - V\gamma d_L \) or \( V = \frac{P_L^{(i)} - P_L^{(2)}}{\gamma(d_L - d_H)}. \) A user with \( v \) should prefer service level \( H \) if \( v \in [V,1] \), or level \( L \) if \( v \in [0,V] \); this is the incentive compatibility constraint (ICC). Even though one level of service may dominate the other, it would be chosen only if the user obtains a non-negative net utility from it. This implies that \( v - P_L^{(i)} - v\gamma d_H \geq 0 \) or \( v \geq \frac{P_L^{(i)}}{1 - \gamma d_H} \); this is the individual rationality constraint (IRC). Let \( V_{L,1} = \frac{P_L^{(2)}}{1 - \gamma d_L} \). Then, combining ICC and IRC, we can express the effective arrival rates from users of segment 1 for the two providers as [14]:

\[
\lambda_{1,1}^{(i)} = \lambda_0(1 - \max\{V, V_{H,1}\}) \quad \text{and} \quad \lambda_{2,1}^{(i)} = \lambda_0(\max\{V, V_{L,1}\} - V_{L,1}).
\]

Also for users of segment 2, let \( V_2 \) be the \( v \)-value of the marginal user who is indifferent between two providers. Hence, we have \( \alpha V_2 - P_L^{(i)} - \alpha V_2\gamma d_H = \alpha V_2 - P_L^{(2)} - \alpha V_2\gamma d_L \) or \( V_2 = \frac{P_L^{(i)} - P_L^{(2)}}{\alpha\gamma(d_L - d_H)}. \) Furthermore, we have \( \alpha v - P_L^{(i)} - \alpha v\gamma d_H \geq 0 \) or \( v = \frac{P_L^{(i)} + P_L^{(2)}}{\alpha(1 - \gamma d_H)} \).

Now, let \( V_{L,2} = \frac{P_L^{(2)} + P_L^{(3)}}{1 - \gamma d_L} \). So, we can express the effective arrival rates from users of segment 2 for the two providers as [14]:

\[
\lambda_{0,2}^{(i)} = \lambda_0(1 - \max\{V_2, V_{H,2}\}) \quad \text{and} \quad \lambda_{2,2}^{(i)} = \lambda_0(\max\{V_2, V_{L,2}\} - V_{L,2}).
\]

Lemma 4: At equilibrium, \( V_{H,1} < V, V_{L,1} < V, V_{H,2} < V_2 \) and \( V_{L,2} < V_2 \).

Proof: First, note that \( V_{L,1} < V \), since otherwise, we have \( \lambda_{2,1} = 0 \), and hence payoff of provider 2 in segment 1 will be equal to zero. So, provider 2 can increase his/her profit by decreasing \( P_L^{(2)} \) till \( V_{L,1} \) drops to a value just below \( V \). because, in that case, \( \lambda_{2,1} > 0 \) and provider 2 enjoys a positive profit from users of segment 1 while satisfying the service level agreement. Now, we can write,

\[
\frac{P_L^{(i)}}{1 - \gamma d_L} < \frac{P_L^{(i)} - P_L^{(2)}}{\gamma(d_L - d_H)} \Rightarrow \frac{P_L^{(i)} - P_L^{(2)}}{P_L^{(i)}} > \frac{\gamma(d_L - d_H)}{1 - \gamma d_L} \Rightarrow \frac{P_L^{(i)} - P_L^{(2)}}{P_L^{(i)}} > \frac{\gamma(d_L - d_H)}{1 - \gamma d_L} + 1 = \frac{1 - \gamma d_H}{1 - \gamma d_L}.
\]

This means that

\[
\frac{P_L^{(i)}}{P_L^{(i)}} < \frac{1 - \gamma d_L}{1 - \gamma d_H} \Rightarrow 1 - \frac{P_L^{(i)} - P_L^{(2)}}{P_L^{(i)}} > 1 - \frac{1 - \gamma d_L}{1 - \gamma d_H} \Rightarrow \frac{P_L^{(i)} - P_L^{(2)}}{P_L^{(i)}} > \frac{\gamma(d_L - d_H)}{1 - \gamma d_H}.
\]

Therefore,

\[
V_{H,1} = \frac{P_L^{(i)}}{1 - \gamma d_H} < \frac{P_L^{(i)} - P_L^{(2)}}{\gamma(d_L - d_H)} = V.
\]

This completes proofs of \( V_{H,1} < V \) and \( V_{L,1} < V \). Using the same argument for users of segment 2, we can prove that \( V_{H,2} < V_2 \) and \( V_{L,2} < V_2 \).

Therefore, the total effective arrival rates for the two providers are:
Lemma 5: When the monopoly (leader) chooses level H and then the two competitors (followers) make a differentiated market, the optimal prices are given by:

\[
\begin{align*}
\lambda_i^{(H)} &= \lambda_{i,1}^{(H)} \lambda_{i,2}^{(H)} = \lambda_{0}(2 - \frac{(\alpha + 1)(P_1^{(H)} - P_2^{(L)})}{\alpha \gamma (d_L - d_H)}) \\
\lambda_2^{(L)} &= \lambda_{2,1}^{(L)} \lambda_{2,2}^{(L)} = \lambda_{0}\frac{(\alpha + 1)(P_1^{(H)} - P_2^{(L)})}{\alpha \gamma (d_L - d_H)} - \frac{(\alpha + 1)P_2^{(L)} + P_3^{(H)}}{\alpha (1 - \gamma d_L)}.
\end{align*}
\]

**Proof:** When one of the competitors chooses service level H and the other L, the providers' optimization problems can be written as:

\[
\begin{align*}
P_1^{(H)} &= \frac{1}{\alpha + 1} \left[ x_1(4\alpha(1 - \gamma d_H) + x_3(x_3 - \alpha - 1)) \quad \text{if} \quad \rho < \rho_r, \right. \\
&\quad + x_4((4\alpha(1 - \gamma d_H) + x_4x_6)(1 - \frac{u_H}{2\rho}) - x_6(\alpha + 1)) \quad \text{if} \quad \rho_r \leq \rho < \rho_s \quad (10) \\
2\alpha(1 - \gamma d_H) - \alpha x_7 - (x_7 + 2\alpha(1 - \gamma d_H))(\frac{u_H}{2\rho}) - (x_7 + 2\alpha(1 - \gamma d_L))(\frac{u_L}{2\rho}) \quad \text{Otherwise,} \\
\end{align*}
\]

\[
\begin{align*}
P_2^{(L)} &= \frac{1}{\alpha + 1} \left[ x_3(2\alpha(1 - \gamma d_L) + 2x_5(x_5 - \alpha - 1)) \quad \text{if} \quad \rho < \rho_r, \right. \\
&\quad + x_4((2\alpha(1 - \gamma d_L) + x_4x_6)(1 - \frac{u_H}{2\rho}) - x_6(\alpha + 1)) \quad \text{if} \quad \rho_r \leq \rho < \rho_s \quad (11) \\
(2(1 - \gamma d_L) - x_7)(\alpha - \frac{u_H + u_L}{2\rho}) \quad \text{Otherwise,} \\
\end{align*}
\]

\[
\begin{align*}
P_3^{(H)} &= \left\{ \begin{array}{ll}
x_5(\alpha + 1 - x_3) \quad & \text{if} \quad \rho < \rho_r, \\
x_6(\alpha + 1 - x_4x_6)(1 - \frac{u_H}{2\rho}) \quad & \text{if} \quad \rho_r \leq \rho < \rho_s \quad (12) \\
x_7(\alpha + 1 - x_5)(1 - \frac{u_H + u_L}{2\rho}) \quad & \text{Otherwise,} \\
\end{array} \right.
\end{align*}
\]

where:

\[
\begin{align*}
x_3 &= \frac{\gamma(d_L - d_H)}{3 - \gamma(4d_H - d_L)}, \\
x_4 &= \frac{\gamma(d_L - d_H)}{1 - \gamma(2d_H - d_L)}, \\
x_5 &= \frac{\alpha(1 - \gamma d_L)(1 - \gamma d_H)}{\alpha(\alpha + 1)(1 - \gamma d_L) + (1 - \gamma d_H)(\alpha + 1 - 2x)}, \\
x_6 &= \frac{x_0 \alpha(1 - \gamma d_H)(1 - \gamma d_H)}{\alpha(\alpha + 1)(1 - \gamma d_L) + (1 - \gamma d_H)(\alpha + 1 - 2x)}, \\
x_7 &= \frac{(1 - \gamma d_H)(1 - \gamma d_H)}{(\alpha + 1)(1 - \gamma d_L) + (1 - \gamma d_H)(\alpha + 1 - 2x)}, \\
x_8 &= \frac{2(2\alpha(1 - \gamma d_L)x_4 - x_3) - (\alpha + 1)(x_4x_6 + 2x_5x_3) + x_4^2x_6 - 2x_5^2x_3}, \\
\end{align*}
\]

and

\[
\begin{align*}
r_r = \frac{u_H(2\alpha(1 - \gamma d_L) + x_4x_6)}{2(2\alpha(1 - \gamma d_L)(1 - \gamma d_L) + x_3x_5 + x_4^2x_6 + u_L(2\alpha(1 - \gamma d_L) + x_5)}, \\
r_s = \frac{u_H(2\alpha(1 - \gamma d_L)(1 - \gamma d_L) + x_4x_6 + u_L(2\alpha(1 - \gamma d_L) + x_5)}{2(2\alpha(1 - \gamma d_L)(1 - \gamma d_L) + x_3x_5 + x_4^2x_6 + x_5 - 2x_5^2x_3}. \\
\end{align*}
\]

Proof: When one of the competitors chooses service level H and the other L, the providers' optimization problems can be written as:

\[
\begin{align*}
\max P_1^{(H)} &= P_1^{(H)} \lambda_0(2 - V - V_2) \quad \text{and} \\
\max P_2^{(L)} &= P_2^{(L)} \lambda_0(V + V_2 - V_L - V_{L,2}) \\
\text{s.t.} \quad b \leq d &\quad \text{and} \\
\frac{1 - \rho(2 - V - V_2)}{1 - \rho(V + V_2 - V_L - V_{L,2})} &\leq d
\end{align*}
\]
Each one is a nonlinear constrained optimization problem with one decision variable and one constraint (due to service-level agreement). If both of the constraints are slack, we can obtain the best response functions from the first order conditions as:

\[ P_1^{(H)} = \frac{\alpha y (d_L - d_H)}{\alpha + 1} + \frac{P_2^{(L)}}{2} \quad \text{and} \quad P_2^{(L)} = \frac{(1 - \gamma d_L) P_1^{(H)}}{2(1 - \gamma d_H)} - \frac{\gamma (d_L - d_H) P_3^{(H)}}{2(\alpha + 1)(1 - \gamma d_H)}. \]

which then yield:

\[ P_2^{(L)} = \frac{2 \alpha (1 - \gamma d_L) x_3}{(\alpha + 1)} - \frac{2 x_3 P_1^{(H)}}{(\alpha + 1)} \quad \text{and} \quad P_1^{(H)} = \frac{4 \alpha (1 - \gamma d_H) x_3 - x_3 P_3^{(H)}}{(\alpha + 1)}. \]

Because of our assumption that the total response time, \( d_L \) and \( d_H \) guarantees, are well separated, if only one of the constraints is binding, it must then be the one for the service with level H; that for the service with level L must be slack. In that case, the best response functions are given by:

\[ P_1^{(H)} = \frac{\alpha y (d_L - d_H)}{\alpha + 1} (2 - \frac{u_H}{\rho}) + P_2^{(L)} \quad \text{and} \quad P_2^{(L)} = \frac{(1 - \gamma d_L) P_1^{(H)}}{2(1 - \gamma d_H)} - \frac{\gamma (d_L - d_H) P_3^{(H)}}{2(\alpha + 1)(1 - \gamma d_H)}. \]

Solving the two equations simultaneously, we get,

\[ P_1^{(H)} = \frac{2 \alpha x_4 (1 - \gamma d_H)}{(\alpha + 1)} (2 - \frac{u_H}{\rho}) - \frac{x_4 P_3^{(H)}}{(\alpha + 1)} \quad \text{and} \quad P_2^{(L)} = \frac{\alpha x_4 (1 - \gamma d_L)}{(\alpha + 1)} (2 - \frac{u_H}{\rho}) - \frac{x_4 P_3^{(H)}}{(\alpha + 1)}. \]

Finally, if both constraints are binding, then the optimal prices can be obtained directly from the constraints as:

\[ P_1^{(H)} = \frac{\alpha (1 - \gamma d_H)}{\alpha + 1} (2 - \frac{u_H}{\rho}) - \frac{\alpha (1 - \gamma d_L)}{\alpha + 1} (\frac{u_L}{\rho} - P_3^{(H)}) \quad \text{and} \quad P_2^{(L)} = \frac{\alpha (1 - \gamma d_L)}{\alpha + 1} (2 - \frac{u_H}{\rho}). \]

Furthermore, in this case, the effective arrival rate for provider 3 from users of segment 3 and segment 2 can be obtained as \( \lambda_3^{(H)} = \lambda_0 (2 - \frac{P_3^{(H)}}{1 - \gamma d_H} - \frac{P_3^{(H)} + P_2^{(L)}}{\alpha (1 - \gamma d_L)}). \) Since there is no capacity limitation for provider 3, the optimization problem can be written as:

\[ \max_{P_3^{(H)}} \pi_3^{(H)} = \lambda_0 P_3^{(H)} (2 - \frac{P_3^{(H)}}{1 - \gamma d_H} - \frac{P_3^{(H)} + P_2^{(L)}}{\alpha (1 - \gamma d_L)}). \]

Now, substituting \( P_2^{(L)} \) from (13) into (14) and using first order condition of \( \pi_3^{(H)} \) we can obtain \( P_3^{(H)} \) in each situation and then \( P_2^{(L)} \) and \( P_1^{(H)} \) will be obtained. In order to complete the proof, the thresholds of \( \rho \) can be found by comparing the optimal prices.

### 4. Equilibrium strategies

The payoff to a provider depends not only on his/her own choice, but also on the choice by his/her competitor and the choice of monopoly provider offering a service that is complementary to their services. Here we assume that the monopoly always offers his/her service with level H. Therefore, if both competitors chose the same service level, they both can charge their users equally as is given by Eqs. (3) and (8). If provider 1 chooses service level H, and provider 2 chooses service level L, then they can charge their users respectively by \( P_1^{(H)} \) and \( P_2^{(L)} \) as given by Eqs. (10) and (11). In each case, payoffs of each provider can be obtained
by multiplying his/her optimal price to his/her total effective arrival rate. Since the functional forms of these profits vary with overall traffic intensity, \( \rho \), different equilibrium strategies could exist.

We have simulated our model as a queuing system. Optimal profits of providers 1, 2 and 3 against \( \rho \) respectively are plotted in figures 3, 4, and 5, when \( b = 0.25s, \alpha = 2, d_m = 0.5s, d_L = 0.8 \) and \( \gamma = 1 \). In the figures, filled markers show the equilibrium strategies.

![Figure 3. Optimal profits of provider 1](image)

![Figure 4. Optimal profits of provider 2](image)

![Figure 5. Optimal profits of provider 3](image)

As seen in figure 3 and 4, if the monopoly chooses level H, when the traffic intensity is low (\( \rho < \rho_C \)), the capacity constraint is not relevant, and the service-level agreement does not have an impact on the providers' decision. Analogous to the traditional setting, see[8], both providers benefit by locating far away from each other in the service dimension, since the competition in that position is the weakest. Consequently, the providers charge different prices for differentiated services.

The figures also show that as the traffic intensity increases above \( \rho_C \), service provider 2 becomes interested in the more lucrative market of service level H. In response, however, provider 1 chooses not to move. Hence, both providers end up in the same market, charging the same price and splitting the demand equally. Even though price competition intensifies in this situation, provider 2 still enjoys a higher profit by switching because of the higher price he/she can charge. The symmetric equilibrium may appear a bit surprising at first, but because of fixed processing capacity, service guarantee and the choice of provider 3 (service level H), it is not surprising.

According to Zhang et al [15] in a duopoly when the traffic intensity is very high, providing the service with level L is beneficial. This is because, under a heavy traffic, the high negative externality imposed by the service level H requires a provider to drastically increase prices (thus resulting in significantly low demand and profit). But in our model because of presence of the monopoly provider who chooses service with level H, when the traffic intensity is very high, providing the service with level H is more beneficial than service with level L.

The optimal demands and optimal prices of the providers can be seen respectively in Figures 6 to 8 and Figures 9 to 11. Similarly, filled markers show the equilibrium states.

![Figure 6. Optimal demands of provider 1](image)

![Figure 7. Optimal demands of provider 2](image)
5. Conclusion

Pricing has been used as an incentive mechanism to control traffic in many areas. As web services become popular, consumers are asking for service level agreements that guarantee the QoS they pay for. In this context, it is particularly important that service providers design an efficient pricing mechanism to enforce service-level agreements. Here, combining game theory and queuing theory, we developed a model to study the duopoly competition where their services are complementary to a service that is offered by a monopoly service provider; service providers can provide either H or L service level. We found out that if the monopoly service provider had chosen service level H, when the traffic intensity was high, service the competitors might choose to compete head to head by providing the same service level H, while, in low traffic intensity both the providers benefit by locating far away from each other in the service dimension and they choose differentiated strategy.

There are several directions for future research. We have assumed that a provider offers only one level of service. A natural extension is to study the duopoly competition if a provider is allowed to provide multiple priority-based service levels. In that case, the queuing model would be more complicated and providers would have more choices. It would be interesting to see how that affects the pricing strategy. Furthermore, in this paper we focused our attention to only one aspect of quality of service, namely the response time. As mentioned earlier, in practice, several aspects of quality are important with respect to web services. We are currently examining how this framework can be extended to multiple dimensions of quality of service.

References


