

Analysis and Comparison of Queues with Different Levels of Delay Information

August 22, 2003; last revised January 7, 2006

Pengfei Guo • Paul Zipkin

Fuqua School of Business, Duke University, Durham, NC 27708, USA

Information about delays can enhance service quality in many industries. Delay information can take many forms, with different degrees of precision. Different levels of information have different effects on customers and so on the overall system. To explore these effects, we consider a queue with balking under three levels of delay information: no information, partial information (the system occupancy) and full information (the exact waiting time). We assume Poisson arrivals, independent, exponential service times, and a single server. Customers decide whether to stay or balk based on their expected waiting costs, conditional on the information provided. We show how to compute the key performance measures in the three systems, obtaining closed-form solutions for special cases. We then compare the three systems. We identify some important cases where more accurate delay information improves performance. In other cases, however, information can actually hurt the provider or the customers.

(Balking Queue; Impatient Customers; Delay Information; Equilibrium Analysis)

1. Introduction

Information flow between customers and providers is an important element of most service systems. Developments in technology and managerial practice can enhance this information flow. In particular, in many situations, it is now possible for the provider to acquire and convey to customers fairly accurate information about anticipated delays due to congestion. Such information can directly affect customer satisfaction and also influence customers' behavior. Here, we focus on the arriving customer's decision to wait for service or to leave (balk).

For example, in a call center, the provider can announce the expected waiting time to each caller; standard call-center software can do this automatically. In transportation and e-shopping, a customer can easily learn the order status and the estimated shipping time. A quote for production services normally includes a lead-time estimate. In a busy hospital emergency room, information about the anticipated wait is important to an anxious patient. Delay notification is used in traffic-flow control; in some cities, congestion conditions are shown on big electronic boards at highway entrances. A similar idea has been suggested for control of computer networks; see, e.g., Kelly (2000).

Information can take many forms, with different degrees of precision. Different levels of information have different effects on customers' decisions and thus the overall arrival process. The goal of this paper is to develop and analyze models to explore these effects. We consider three typical types of delay information: none, the system occupancy, and the exact waiting time. With *no information*, customers still estimate their waiting times, but these estimates are based only on long-term (equilibrium) experience, not real-time information. The occupancy provides *partial information*; the remaining uncertainty comprises the actual service times of the waiting customers. The exact waiting time gives the customer *full information*. We assume that the information provided is truthful, and customers believe it.

Depending on the situation, it can be hard or easy, expensive or cheap, for the provider to obtain delay information and transmit it to customers. Some amusement parks post signs in waiting areas indicating expected delays based on distance to the head of the line. That information is cheap. In other settings, information requires a substantial investment in technology. Here, we do not discuss such costs, but rather focus on the benefits.

We assume Poisson arrivals and independent, exponential service times. There is a single server using the FCFS discipline.

To assess the effects of information consistently, we posit a customer-decision mechanism common to all levels of information. A given function measures the *basic* cost of delay. Different customers, however, value time differently. Each customer arrives with a specific parameter, which scales the basic cost function. Upon arrival the customer receives information, which affects his estimate of the distribution of delay. Based on his scale parameter and the information, the customer computes his *expected* delay cost. If that is more than the reward he anticipates from receiving service, he balks, and if not, he stays. In this scheme, then, different levels of information lead to different delay distributions in the expected-cost calculations, and those in turn affect everything else. We assume that the service provider cares about the throughput, perhaps because he receives a payment for each customer served.

We show how to compute the key performance characteristics for each of the three information scenarios. Then, we compare the results to assess the impacts of information. These impacts turn out to be subtle. One might expect more information always to be better, but it's not. The effect depends mainly on the form of the cost-scale parameter distribution.

An overview of customer psychology in waiting situations, including the impact of uncertainty, can be found in Maister (1984). There is some empirical evidence about customers' reactions to delays. Taylor (1994) shows that delays affect customers' overall service evaluations. Hui and Tse (1996), Kumar et al. (1997) and Munichor and Rafaeli (2005) study the relationship between information and customer satisfaction. Zhou and Soman (2003) examine the determinants of renegeing behavior.

There is a substantial literature on queues with impatient customers. Models with balking and renegeing (leaving after waiting for some time) can be found in many books, e.g. Kulkarni (1995). Recent works on this topic include Bae, Kim and Lee (2001), Zohar et al. (2002), and Ward and Glynn (2003).

The literature on customers influenced by delay information begins with Naor (1969), who studies a system like ours with partial information, but with identical customers and linear waiting cost. Also, the cost depends on the whole sojourn time, not just the delay. He points out that this system with its self-selecting customers suffers from *externalities*; an arriving customer who stays imposes delays on later customers, but ignores them in making his decision. Consequently, too many customers stay. If everyone were altruistic and acted to maximize the average utility, some of those customers would leave. He shows that a price can steer the system to this "social" optimum. He points out, however, that if the price is determined by the provider to maximize revenue, the provider becomes a monopolist and

behaves like one. He sets the price higher than the socially optimal one and thus serves too few customers.

These features lead to additional peculiarities. Edelson and Hildebrandt (1975) mention that a revenue-maximizing service provider may make socially wrong decisions about service capacity, either in the service rate or the number of servers. They also show that the difference between the monopoly solution and the socially optimum is due entirely to balking. Schroeter (1982) considers non-identical customers with uniformly distributed costs. Hassin and Haviv (2003) summarize additional research along these lines. Gavish and Schweitzer (1973) study a full-information system under similar assumptions.

More recently, Mandelbaum and Shimkin (2000) and Shimkin and Mandelbaum (2004) discuss equilibria with respect to renegeing in a setting similar to our no-information model, with linear and nonlinear waiting cost functions, respectively. Afèche and Mendelson (2004) study revenue-maximizing and socially optimal equilibria under uniform pricing with no information. The paper also discusses priority auctions. Whitt (1999) studies two systems corresponding roughly to our models with no and partial information. His customer-choice mechanism, however, is quite different from ours, and so are his findings about the impact of information on performance. There, information always reduces both waiting and throughput. Thus, despite their similar motivation, his models represent very different behavior from ours.

Armony and Maglaras (2004a, 2004b) analyze systems where an arriving customer learns some delay information and then can choose to balk, wait, or leave a message, in which case the provider calls back within a guaranteed time. That guaranteed time is an estimate based on a heavy-traffic approximation. In that heavy-traffic regime, indeed, such estimates become nearly precise. In Armony and Maglaras (2004a) customers' choices are based on the equilibrium waiting time, as in our no-information system. Customers employ a utility function to assess delays, but the function's argument is the expected delay, a constant. This approach is justified in the heavy-traffic limit, but it thereby suppresses the risk-reduction role of information. In Armony and Maglaras (2004b) each arriving customer receives more information, a point estimate of delay based on the system occupancy. The customer treats the estimate as exact, again based on the heavy-traffic limit. Comparing the results with the other system, they show that more information improves performance on several dimensions. Thus, their modeling approach is quite different from ours, and their findings support a more optimistic view of the role of information.

Another stream of research explores lead-time quotation in production, e.g., Duenyas and Hopp (1995) and Spearman and Zhang (1999). The system studied by Dobson and Pinker (2000) is related to ours. The provider quotes a number, the nominal lead-time, to every arriving customer. This is understood by all parties to mean a certain fixed fractile of the lead-time distribution. The provider himself may use different levels of information to assess this distribution and hence the fractile, but customers see only the nominal leadtime. Their responses are determined by a demand function – more customers stay when the quoted lead-time is shorter. The paper shows that the impact of more information depends on the shape of this demand function, a notion related to our findings about the cost-scale distribution. It only examines the impact on the provider, however, not the customers.

Social optimization aims to achieve the best outcome for everyone. It usually requires some central coordination scheme, such as admission control. There, more information is always better than less (absent information-processing costs), regardless of the objective. The controller can choose to ignore additional information, so the less-information solution is feasible for the more-information case. In our system, with its individual decisions, the matter is less obvious. We show for some important cases that more information does improve performance, for the provider or the customers or both, but for other cases it does not.

The remainder of the paper is organized as follows: Section 2 introduces the basic formulation. Sections 3-5 develop the models for no information, partial information, and full information, respectively. Section 6 compares the three systems analytically. Section 7 treats an important extension. Section 8 presents some concluding remarks. A supplement contains proofs and other technical material.

2. Formulation and Preliminaries

Potential customers arrive according to a Poisson process with rate λ . There is a single server, and the service times are independent and exponentially distributed with mean $1/\mu$. The system uses the FCFS discipline.

2.1 Customer Behavior

We suppose that a customer's utility equals a reward for receiving service minus a waiting cost. This waiting cost depends on a customer-specific parameter and the expectation of a

common function of the waiting time. (In some applications, such as production, the total sojourn time is more important than the delay. The approach can be extended to that case.) Specifically, define

- W = waiting time in queue
- θ = customer-type parameter, indicating the importance of time, $\theta \in [0, 1]$
- H = cumulative distribution function of θ , continuous on $[0, 1]$, with density h
- $c(w)$ = basic cost to wait time w , a positive, increasing, unbounded, continuous function
- r = reward to the customer for receiving service, $r > 0$.

The service reward r is the same for all customers (this assumption is convenient but not essential). Customers differ in the importance of time. This difference is expressed by the customer type θ . Each customer's type is independent of all other events. The customer assesses the distribution of waiting time W , based on the available information. The expected waiting cost $\theta E[c(W)]$ is a function of that information. The utility U for the customer to stay is then

$$U = r - \theta E[c(W)].$$

The customer remains in the system if U is non-negative and otherwise balks.

We can rescale r and c so that $r = 1$. Suppose $c(0) > 1$. Then, some potential customers, precisely those with $\theta > 1/c(0)$, always balk. We can simply ignore them and scale down λ and c to represent the other customers. Thus, we can assume $c(0) \leq 1$. For convenience, we assume for now that $c(0) = 1$. Thus, the reward is just large enough to attract the most sensitive customers when there is no delay. We defer the case $c(0) < 1$ to Section 7.

Different cost functions express different sensitivities to risk, just like utility functions for wealth in finance and economics. A strictly convex cost means a strong aversion to risk, while a linear cost expresses indifference to risk. On the other hand, it is easy to think of situations where the marginal cost of waiting decreases, so the cost is concave. (“We’ve already waited a whole hour, a few more minutes won’t make any difference.”)

2.2 Average Utility and Throughput

Let I stand for information, a random variable with possible values i . Given information $I = i$, an arriving customer computes the expected waiting cost, which we can write $E_W[c(W)|I = i]$. The customer stays, then, precisely when his θ is less than or equal to the critical level $\theta_i = 1/E_W[c(W)|I = i]$. In these terms, the overall fraction of customers who stay is $E_I[H(\theta_I)]$, the throughput is $\lambda E_I[H(\theta_I)]$, and the probability the server is busy is $(\lambda/\mu)E_I[H(\theta_I)]$.

Define

$$J(\theta) = (1/\theta) \int_0^\theta H(\phi) d\phi.$$

The average expected utility for customers is

$$\begin{aligned} u &= E[U^+] = E_{\theta, I} [[1 - \theta E_W[c(W)|I]]^+] \\ &= E_I \left[\int_0^{\theta_I} (1 - \phi E_W[c(W)|I]) h(\phi) d\phi \right] \\ &= E_I \left[H(\theta_I) - (1/\theta_I) \int_0^{\theta_I} \phi h(\phi) d\phi \right] \\ &= E_I[J(\theta_I)]. \end{aligned}$$

3. No Information

First consider the situation where the queue is invisible, and the provider tells the customer nothing. Suppose W is the equilibrium waiting time, and all customers know its distribution. An arriving customer has two choices, stay or balk. The customers who stay are precisely those with $\theta \leq \theta_-$, where $\theta_- = 1/E[c(W)]$. Consequently, the fraction of customers who stay is $H(\theta_-)$, and the effective arrival process is Poisson with rate $\lambda_- = \lambda H(\theta_-)$. This effective arrival rate affects W , hence $E[c(W)]$, and hence θ_- . We assume that these parameters arrive at consistent, i.e., equilibrium values. (For discussions of such equilibria in related models, see Stidham 1985 and Hassin and Haviv 2003.)

In sum, the equilibrium arrival rate λ_- solves

$$\lambda_- = \lambda H \left(\frac{1}{E[c(W| -)]} \right). \quad (1)$$

Here, $E[c(W| -)]$ indicates the expected cost given λ_- . Assume that it is finite for any $\lambda_- < \mu$.

Proposition 1 *For no information, there exists a unique equilibrium arrival rate λ_- .*

(The proof of this and the other results are in the supplement.) The assumption of finite $E[c(W| -)]$ is necessary. Consider $c(w) = e^{\beta w}$ with $\beta \geq \mu$ and exponential service times. For any $\lambda_- > 0$, $E[c(W| -)] = \infty$, while for $\lambda_- = 0$, $E[c(W| -)] = 1$. So, there is no equilibrium. In words, if nobody comes, everybody wants to come, but if anybody comes, nobody wants to come.

Since all customers receive the same information (none), the system behaves as an $M/M/1$ queue. Thus, W has the truncated exponential distribution with rate $\mu(1 - \rho_-)$ and mass $1 - \rho_-$ at 0, where $\rho_- = \lambda_-/\mu$. Consequently,

$$E[c(W| -)] = (1 - \rho_-) + \rho_- \{\mu(1 - \rho_-)\tilde{c}[\mu(1 - \rho_-)]\},$$

where \tilde{c} denotes the Laplace transform of c . Thus, the equilibrium ρ solves

$$\rho = (\lambda/\mu)H \left(\frac{1}{(1 - \rho) + \rho \{\mu(1 - \rho)\tilde{c}[\mu(1 - \rho)]\}} \right). \quad (2)$$

The average utility is then $J(\theta_-)$, where $\theta_- = 1/E[c(W| -)]$.

Example 1: Uniform customers with linear cost

Suppose H is the uniform distribution, and $c(w) = 1 + w$. Then, $\tilde{c}(s) = 1/s + 1/s^2$, and (2) becomes

$$\rho = \frac{\lambda/\mu}{1 + \rho/[\mu(1 - \rho)]},$$

or

$$(1 - \mu)\rho^2 + (\mu + \lambda)\rho - \lambda = 0, \quad (3)$$

a quadratic equation. For $\mu = 1$, the root of this equation is

$$\rho = \frac{\lambda}{1 + \lambda}.$$

For $\mu \neq 1$, the positive root less than 1 is

$$\rho = \frac{-(\mu + \lambda) + \sqrt{(\mu + \lambda)^2 + 4\lambda(1 - \mu)}}{2(1 - \mu)}.$$

Similarly, for quadratic cost $c(w) = 1 + w^2$, (2) becomes a cubic equation with a single root between 0 and 1.

4. Partial Information

Suppose the provider observes and tells the customer $N(t)$, the system occupancy at the moment of arrival. The customer computes $c_n = E[c(W)|N(t) = n]$ and elects to stay if $\theta \leq \theta_n$, where $\theta_n = 1/c_n$. So, given $N(t) = n$, the effective arrival process is Poisson with rate $\lambda_n = \lambda H(\theta_n)$. Thus, $N(t)$ is a birth-death process; the birth rate in state n is $\lambda_n = \lambda H(\theta_n)$, and the death rate is μ . Let N denote the equilibrium occupancy and $p_n = \Pr\{N = n\}$. Then, by the standard analysis of birth-death processes,

$$p_n = \left(\prod_{m=0}^{n-1} \lambda_m / \mu \right) p_0 = \Theta_n (\lambda / \mu)^n p_0,$$

where

$$\Theta_n = \prod_{m=0}^{n-1} H(\theta_m), \quad n > 0.$$

Let

$$\Theta = \sum_{n>0} \Theta_n (\lambda / \mu)^n.$$

Then,

$$p_0 = \frac{1}{1 + \Theta}.$$

Note that Θ is finite, because $\theta_n \rightarrow 0$. The customers' decisions ensure a stable system, even for cases like $c(w) = e^{\beta w}$, where the no-information system fails to reach equilibrium.

Example 2: Uniform customers with linear cost

$$\begin{aligned} \Theta_n (\lambda / \mu)^n &= \left(\prod_{m=0}^{n-1} \frac{1}{1 + m / \mu} \right) (\lambda / \mu)^n \\ &= \frac{\lambda^n}{\mu(\mu + 1) \cdots (\mu + n - 1)} \\ &= \frac{\Gamma(\mu)}{\Gamma(\mu + n)} \lambda^n \\ \Theta &= \sum_{n=1}^{\infty} \frac{\Gamma(\mu)}{\Gamma(\mu + n)} \lambda^n = \lambda^{1-\mu} e^{\lambda} \gamma(\mu, \lambda) \end{aligned}$$

$$\begin{aligned} p_0 &= \frac{1}{1 + \gamma(\mu, \lambda) \lambda^{1-\mu} e^{\lambda}} \\ p_n &= p_0 \frac{\Gamma(\mu)}{\Gamma(\mu + n)} \lambda^n, \quad n > 0, \end{aligned}$$

where $\Gamma(\mu) = \int_0^\infty t^{\mu-1} e^{-t} dt$ is the gamma function, and $\gamma(\mu, \lambda) = \int_0^\lambda t^{\mu-1} e^{-t} dt$ is the lower incomplete gamma function; see Abramowitz and Stegun (1965). We have obtained the distribution of N in closed form. This is a generalization of the Poisson distribution. The Poisson is the special case with $\mu = 1$. (Ward and Glynn 2003 derive a similar distribution for a different system, one with reneging.) The expected occupancy can be expressed in a simple form (see the supplement):

$$E[N] = \lambda - (\mu - 1)(1 - p_0). \quad (4)$$

5. Full Information

Now, suppose the provider observes and tells each arriving customer the exact waiting time. (For instance, each arriving customer brings the realization of his service time. This is nearly true in some production systems.) So W is a constant for the customer. The workload (or virtual waiting time) $V(t)$ is the total time needed to complete the service of all customers currently in the system. Then, the waiting time for an incoming customer at t equals $V(t)$. The effective arrival rate given workload v is $\lambda(v) = \lambda H(\theta_v)$, where $\theta_v = 1/c(v)$. The sample path of $V(t)$ works in the usual way: When $V(t) > 0$, it decreases at constant rate -1 ; when a customer joins the system, $V(t)$ increases by the service time of that customer.

Denote the density function of the equilibrium workload V by $f(v)$, $v > 0$, and let p_0 be its mass at 0. By a level-crossing argument (Brill and Posner 1977), these quantities uniquely satisfy the integral equation

$$f(v) = \lambda p_0 e^{-\mu v} + \int_0^v \lambda H[1/c(w)] e^{-\mu(v-w)} f(w) dw \quad (5)$$

and the normalization condition

$$p_0 + \int_0^\infty f(v) dv = 1. \quad (6)$$

One can easily check that the solution is as follows: Define

$$C(v) = \int_0^v H[1/c(t)] dt.$$

Then,

$$f(v) = \lambda p_0 e^{\lambda C(v) - \mu v},$$

where

$$p_0 = \frac{1}{1 + \lambda \int_0^\infty e^{\lambda C(v) - \mu v} dv}.$$

We now have the solution in closed form, up to the evaluation of these integrals. (It is not hard to show that the integrals are finite. Again, the customers' decisions ensure stability.)

Example 3: Uniform customers with linear cost

$$C(v) = \int_0^v \frac{1}{1+t} dt = \ln(1+v).$$

Thus,

$$\begin{aligned} \int_0^\infty e^{\lambda C(v) - \mu v} dv &= \int_0^\infty (1+v)^\lambda e^{-\mu v} dv \\ &= e^\mu \mu^{-(\lambda+1)} \int_\mu^\infty y^\lambda e^{-y} dy \\ &= e^\mu \mu^{-(\lambda+1)} \Gamma(\lambda+1, \mu), \end{aligned}$$

where $\Gamma(\lambda+1, \mu) = \int_\mu^\infty y^\lambda e^{-y} dy$ is the upper incomplete gamma function; see Abramowitz and Stegun (1965). So,

$$\begin{aligned} p_0 &= \frac{1}{1 + \lambda e^\mu \mu^{-(\lambda+1)} \Gamma(\lambda+1, \mu)}, \\ f(v) &= \lambda p_0 (1+v)^\lambda e^{-\mu v}, \quad v > 0. \end{aligned}$$

This is a truncated gamma distribution.

6. Comparisons

6.1 Cost-Scale Distributions

We first identify some important classes of customer cost-scale distributions H .

Consider a *power distribution* $H(\theta) = \theta^\alpha$ for $\alpha > 0$. (The uniform distribution is the case with $\alpha = 1$.) Here,

$$J(\theta) = \frac{1}{\alpha+1} \theta^\alpha.$$

Thus, J is proportional to H . Consequently, *the average utility* $u = E[J(\theta_I)]$ *is proportional to the throughput* $\lambda E[H(\theta_I)]$. (Clearly, these are the only distributions with this property.)

This is a striking result. One simple performance measure, the busy probability or throughput, serves to characterize *both* the server's profit and the customers' average utility. The provider's and the customers' objectives are perfectly aligned. Put another way, one

need not separately measure customer satisfaction. Just count the money. As we'll see shortly, more information is better for all parties in this case.

This is not always so, however. To see that *some* condition is needed, consider the case of identical customers with linear cost. For sufficiently small λ , under no information, all customers choose to stay, so the throughput is λ . Specifically, this happens when

$$\theta \left(1 + \frac{1}{\mu - \lambda} - \frac{1}{\mu} \right) \leq 1,$$

or

$$\lambda \leq \mu \left(1 - \frac{1}{1 + \mu[(1/\theta) - 1]} \right).$$

Under partial information, however, there is a cutoff point \bar{n} , such that all customers stay when $N \leq \bar{n}$, but they all balk when $N > \bar{n}$. Thus, the throughput is $< \lambda$. Here, information *reduces* throughput. If the queue is readily visible, the provider may try to hide it. (This point is implicit in Hassin 1986.) The same thing happens under full information.

Information can also hurt customers, due to externalities. Suppose there are two types of customers, A and B . Customers of each type are identical, but the types have different θ 's, say $\theta_B \ll \theta_A$. An arriving customer is of type T with probability h_T . Suppose that, in the no-information system, all B customers stay, while all A customers leave. In the partial-information system, some of those A 's stay, when they encounter an empty or near-empty system. They get only slightly positive utility, but they take it. So, the B 's suffer lower utilities, and the loss may overwhelm the benefit to the A 's. Here, the customers may prefer to hide the queue. To illustrate, consider the following case: $\theta_B = 1/8$, $\theta_A = 63/64$, $h_A = h_B = 1/2$, $\lambda = \mu = 1$. No information yields average utility 0.375, while partial information yields 0.358.

To understand when such behavior occurs, we distinguish two broad classes of distributions H . As we'll see, the following condition ensures that more information increases throughput.

Condition 1 *The function $H(1/x)$ is convex in $x \geq 1$.*

This means that the cost-scale distribution is spread out, so customers are heterogeneous, in a certain sense. It is equivalent to

$$-\frac{\theta h'(\theta)}{h(\theta)} \leq 2$$

(assuming the derivative h' exists). The left-hand side is the elasticity of the density h . The condition posits that h not be too elastic, that is, that customers not be too concentrated. More precisely, it rules out a sharp *decrease* in h . It is a one-sided spread condition.

For example, consider the beta density

$$h(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

with parameters $\alpha, \beta > 0$, where $B(\alpha, \beta)$ is the beta function. This satisfies the condition, if and only if $\beta \leq 1$. The customers are too concentrated if $\beta > 1$.

Next, assume that H is strictly increasing, so it has a well-defined inverse H^{-1} . The following condition guarantees that more information benefits customers.

Condition 2 *The function $J \circ H^{-1}$ is convex on $[0, 1]$.*

One can show that this condition is equivalent to

$$-\frac{\phi h'(\phi)}{h(\phi)} \geq 2 - \frac{\phi h(\phi)}{[H(\phi) - J(\phi)]}.$$

Again, we have a restriction on the elasticity of h , but here it's a *lower* bound, a variable one. It requires that h not *rise* too sharply. For a beta distribution, the condition holds precisely for $\beta \geq 1$.

Note that a power distribution is a beta distribution with $\beta = 1$. Among the beta distributions, these are the only ones that satisfy both conditions.

The following tables show that, even for smooth beta distributions, when these conditions are violated, more information can degrade performance. Table 1 shows the busy probabilities for systems with linear cost ($c(w) = 1 + w$). Here, $\beta > 1$, so Condition 1 is violated. The **bold** numbers indicate where more information hurts the provider. Table 2 displays average utilities in the same format. These cases have $\beta < 1$, so they violate Condition 2. (Systems with nonlinear costs display similar behavior.)

Observe that information hurts the provider only for quite large β , so the customers are quite concentrated, and for light traffic (small λ). Likewise, information hurts the customers only for very small β and large λ . Even in such cases, the effects are small.

6.2 No Information and Partial Information

Let us use superscripts *no*, *part* and *full* to indicate the different information levels. We now compare the performance of the no-information and partial-information systems.

Table 1: Busy Probability with Linear Cost Function
 $\alpha=2, \mu=2$

λ	$\beta=4$			$\beta=8$			$\beta=16$		
	no	partial	full	no	partial	full	no	partial	full
0.5	0.2495	0.2448	0.2439	0.2500	0.2494	0.2486	0.2500	0.2499	0.2498
1	0.4812	0.4677	0.4678	0.4995	0.4918	0.4894	0.5000	0.4990	0.4981
2	0.7216	0.7876	0.8028	0.8179	0.8621	0.8683	0.8871	0.9145	0.9168
4	0.8404	0.9802	0.9925	0.9095	0.9984	0.9997	0.9515	1.0000	1.0000
8	0.9003	0.9998	1.0000	0.9456	1.0000	1.0000	0.9714	1.0000	1.0000

Table 2: Average Utility with Linear Cost Function
 $\alpha=2, \mu=2$

λ	$\beta=0.15$			$\beta=0.1$			$\beta=0.05$		
	no	partial	full	no	partial	full	no	partial	full
0.5	0.0570	0.0594	0.0612	0.0400	0.0405	0.0418	0.0215	0.0207	0.0214
1	0.0511	0.0523	0.0552	0.0365	0.0356	0.0377	0.0201	0.0182	0.0193
2	0.0436	0.0431	0.0472	0.0319	0.0294	0.0323	0.0182	0.0151	0.0167
4	0.0351	0.0335	0.0384	0.0265	0.0230	0.0266	0.0159	0.0119	0.0139
8	0.0262	0.0251	0.0301	0.0206	0.0176	0.0215	0.0131	0.0092	0.0115

Proposition 2 If $p_0^{part} \geq p_0^{no}$, then $u^{part} > u^{no}$.

Thus, more information helps *someone* – if not the provider, then the customers.

Proposition 3 Under Condition 1 [$H(1/x)$ is convex], $p_0^{part} \leq p_0^{no}$.

Thus, for certain shapes of the cost-scale distribution H , more information increases throughput and so helps the provider. For power H , therefore, information also increases customers' average utility.

To compare utilities more generally, we need a different condition on the shape of H .

Proposition 4 Under Condition 2 [$J \circ H^{-1}$ is convex], $u^{part} > u^{no}$. Moreover, if $p_0^{part} < p_0^{no}$, then

$$\frac{u^{part}}{u^{no}} \geq \frac{1 - p_0^{part}}{1 - p_0^{no}}.$$

For such customer distributions, then, average utility improves with information. The throughput may or may not increase. If it does increase, then utility increases even more, proportionally.

Finally, we mention an interesting fact about the special case considered in the examples above.

Proposition 5 *For uniform H and linear cost, the relation between $E[N^{no}]$ and $E[N^{part}]$ is the same as that between μ and 1. That is, they are equal for $\mu = 1$, $E[N^{no}] > E[N^{part}]$ for $\mu > 1$, and $E[N^{no}] < E[N^{part}]$ for $\mu < 1$.*

We have already seen that throughput and utility both increase with more information in this case. However, the standard performance measures such as $E[N]$ need not improve. But such measures are not the most relevant ones in this context. The customers here place different weights on delays. Information allows them to filter themselves, so that those who care more about delays wait less.

6.3 No Information and Full Information

The systems with no and full information are related in exactly the same ways. For example, if $p_0^{full} \geq p_0^{no}$, then $u^{full} > u^{no}$.

6.4 Partial Information and Full Information

It is harder to compare the partial- and full-information systems. We have only the following result:

Proposition 6 *Under Condition 1 [$H(1/x)$ is convex], $p_0^{full} \leq p_0^{part}$.*

Thus, again, with this shape of H , more information increases throughput. And, in case H is a power distribution, the full-information system has higher utility.

We *conjecture* that the other results above, comparing no and partial information, also describe the relation between partial and full information.

7. Extension

Now, suppose $c(0) < 1$. Some of the results above remain valid in this case, but others don't.

Some θ_i are now greater than 1. Let us extend the domain of H and J to all $\theta \geq 0$, setting $H(\theta) = 1$ for $\theta \geq 1$. The definition of J in terms of H remains the same. With this understanding, the throughput is still $\lambda E_I[H(\theta_I)]$, and the average utility is still $E_I[J(\theta_I)]$. The solutions for the three information models remain the same.

Turning to comparisons, let us focus on the relation between the no- and partial-information systems. The other comparisons are similar.

It is no longer true that, for a power distribution H , average utility is proportional to throughput. The incentives of the provider and the customers are *not* perfectly aligned.

Propositions 2 and 4 hold as stated. Thus, more information always helps someone, and under Condition 2, it helps customers.

Proposition 3 is no longer valid. The situation here is identical to the case of identical customers, discussed in §6.1 above. For *any* H , and sufficiently small λ , all customers stay in the no-information system. But, with partial information, some customers balk, so the throughput is lower.

We can obtain a qualified version of the result, however: Under Condition 1, for sufficiently large λ , $p_0^{part} \leq p_0^{no}$. In fact, this holds even under a weaker condition on H , namely, $H(1/x)$ is convex for sufficiently large x . (This covers all beta distributions, even those with $\beta > 1$.) Thus, under this condition, more information may hurt the provider with light traffic, but not with heavy traffic.

8. Conclusions

We considered service systems with three levels of customer-delay information. Customers use that information to determine their expected waiting costs, and so to decide whether to stay and receive service or leave (balk). We obtained closed-form solutions for some cases and nearly closed-form solutions for others. In comparing these systems, we found that the form of the cost-scale distribution plays a crucial role. For one important class, average utility is proportional to throughput; so the provider's and customers' objectives coincide; those measures improve as information increases. More broadly, we found sufficient conditions to ensure that more information helps the provider or the customers. In other cases, however, more information can actually hurt one or the other.

These perverse phenomena occur mainly in extreme conditions, however. Information can make a bad system worse. A larger model, we suspect, would avoid such situations, by adjusting the capacity to the workload or managing demand more actively.

This utility-based approach forces us to revise our notions of good performance. Most peculiar is the concept that customers actually may prefer one system to another when its probability of delay and average delay are larger. Of course, this does not mean that the customers want to wait. Rather, it shows that these standard measures do not capture everything that matters to customers.

Numerous extensions are worth pursuing, for example, general service times, alternative queue disciplines, inventory, and other models of information. It would be interesting also to explore various pricing schemes, following the lead of Naor.

Information can modulate waiting costs in subtler ways than our model envisions. For example, given a delay estimate, a call-center customer may turn attention to other tasks while waiting. In general, information affects people's expectations, and those expectations affect the overall experience of waiting. (Carmon et al. 1995 pose a framework for such effects.) It would be interesting to study how delay information is acquired and used in various situations and the resulting effects on overall system behavior.

References

- Abramowitz, M. and I. Stegun. 1965. *Handbook of Mathematical Functions*. Dover. New York.
- Afèche, P. and H. Mendelson. 2004. Pricing and priority auction in queueing systems with a generalized delay cost structure. *Management Sci.* 50 869-882.
- Armony, M. and C. Maglaras. 2004a. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* 52 271-292.
- Armony, M. and C. Maglaras. 2004b. Contact center with a call-back option and real-time delay information. *Oper. Res.* 52 527-545.
- Bae, J., S. Kim and E. Lee. 2001. The virtual waiting time of the M/G/1 queue with impatient customers. *Queueing Systems.* 38 485-489.
- Brill, P. and M. Posner. 1977. Level crossings in point process applied to queues: single-server case. *Oper. Res.* 25 662-674.
- Carmon, Z., J. Shanthikumar and T. Carmon. 1995. A psychological perspective on service segmentation models: The significance of accounting for consumers' perceptions of waiting and service. *Management Sci.* 41 1806-1815.
- Dobson, G. and J. Pinker. 2000. The value of sharing lead-time information in custom production. Simon Business School Working Paper No. CIS 00-02.
- Duenyas, I. and W. Hopp. 1995. Quoting lead times. *Management. Sci.* 41 43-57.
- Edelson, M. and K. Hildebrand. 1975. Congestion tolls for Poisson queueing processes.

- Econometrica*. 43 81-92.
- Gavish, B. and P. Schweitzer. 1973. Queue regulation policies using full information. Working paper, Israel Scientific Center.
- Hassin, R. 1986. Consumer information in markets with random products quality: The case of queues and balking. *Econometrica*. 54 1185-1195.
- Hassin, R. and M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queuing Systems*. Kluwer. Boston/Dordrecht/London.
- Hui, M. and D. Tse. 1996. What to tell customer in waits of different lengths: an integrative model of service evaluation. *J. Marketing*. 60 81-90.
- Kelly, F. 2000. Models for self-managed Internet. *Phi. Trans. R. Soc. Lond. A* 358 2335-2348.
- Kulkarni, V. 1995. *Modeling and Analysis of Stochastic Systems*. Chapman & Hall.
- Kumar, P., M. Kalwani and M. Dada. 1997. The impact of waiting time guarantees on customers' waiting experiences. *Marketing Science*. 16 295-314.
- Maister, D. 1984. Psychology of waiting lines. *Harvard Business School Cases*. Apr 01, 71-78.
- Mandelbaum, A. and N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems*. 36 141-173.
- Munichor, N. and A. Rafaeli. 2005. Numbers or apologies? customer reactions to tele-waiting time fillers. Technion. Working Paper.
- Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* . 37 15-24.
- Schroeter, R. 1982. The costs of concealing the customer queue. Working paper. Department of Economics. Iowa State University.
- Shimkin, N. and A. Mandelbaum. 2004. Rational abandonment from tele-queues: nonlinear waiting cost with heterogeneous preferences. *Queueing Systems*. 47 117-146.
- Spearman, M. and R. Zhang. 1999. Optimal lead time policies. *Management Sci.* 45 290-295.
- Stidham, S. 1985. Optimal control of admission to a queuing system. *IEEE Trans. Auto. Control*. 30 705-13.
- Taylor, S. 1994. Waiting for service: The relationship between delays and evaluations of

- service. *J. Marketing.* 58 56-69.
- Ward, A. and P. Glynn. 2003. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, 43 103-128.
- Whitt, W. 1999. Improving service by informing customers about anticipated delays. *Management Sci.* 45 192-207.
- Zhou, R. and D. Soman. 2003. Looking back: Exploiting the psychology of queueing and the effect of the number of people behind. *J. Consumer Research.* 29 517-529.
- Zohar, E., A. Mandelbaum and N. Shimkin. 2002. Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Sci.* 48 566-583.