Queueing Systems manuscript No.
(will be inserted by the editor)

Sharing Delay Information in Service Systems: A Literature Survey

Rouba Ibrahim

Received: date / Accepted: date

Abstract Service providers routinely share information about upcoming waiting times with their customers, through delay announcements. The need to effectively manage the provision of these announcements has led to a substantial growth in the body of literature which is devoted to that topic. In this survey paper, we systematically review the relevant literature, summarize some of its key ideas and findings, describe the main challenges that the different approaches to the problem entail, and formulate research directions that would be interesting to consider in future work.

Key words: delay announcements; queueing models; service systems.

1 Introduction

Most service providers routinely share various levels of information with their customers. This information may be about the quality of the service provided, the service design, the current number of people in the system, the duration of anticipated delays, etc. In broad terms, sharing information with customers is deemed beneficial for two main reasons. First, customers usually appreciate receiving some feedback from service providers, rather than being "kept in the dark" during their service experiences; i.e., the quality of service is perceived to be higher because of such feedback. Second, because customers in service systems are people who typically change their behaviors in response to any information that they receive, information-sharing may be viewed as a lever of control in the system which, if managed in an effective manner, could be used to the benefit of the service provider.

Here, we focus on a setting where the service provider shares information about upcoming waiting times with customers, in the form of delay announce-

Rouba Ibrahim
School of Management, University College London
E-mail: rouba.ibrahim@ucl.ac.uk

ments. Nowadays, sharing delay information with customers is widespread: It is used in various service contexts such as amusement parks, call centers, hospitals, retail stores, immigration and border controls, transportation networks, etc. By and large, the announcements play a dual role: First, they increase customer satisfaction by reducing the uncertainty about upcoming waits; and second, they are a tool of voluntary demand modulation because they encourage customers to join low-congestion states, and deter them from joining high-congestion states, i.e., they allow for a better match between demand and supply. Due to both its practical relevance and its theoretical appeal, the problem of effectively managing the provision of delay announcements in service systems is interesting to practitioners and academics alike. In particular, recent years have witnessed a remarkable growth of the body of literature devoted to studying that problem. In this paper, we survey that body of literature.

1.1 Do Customers Appreciate Delay Announcements?

For motivation, we begin by summarizing some of the main findings on the psychological impact of delay announcements. An important maxim in service science is that customers do not like the uncertainty associated with waiting. This finding has been confirmed with airline delays (Taylor [79]), banks (Katz et al. [58]), and websites (Weinberg [84]). Aversion to the uncertainty in waiting is also underlined as one of the axioms in Maister [61]. More generally, Leclerc et al. [59] provide empirical evidence (via experimental study) that waiting may be viewed as a cost for delayed individuals. Delay announcements are useful because they are means to reducing that undesirable uncertainty.

Another psychological benefit of delay announcements relates to the distinction between perceived time and actual time (Hornik [44]). To be specific, the relationship between the perception of time and the evaluation of the waiting experience is mediated by several factors, including the perceived control over time (McGuire et al. [64]). Delay announcements are beneficial because they enable customers to have increased control over their waits, e.g., if the waiting time is sufficiently long, a customer may elect to perform other tasks while waiting. Thus, customer waits may be perceived to be shorter. Even in settings where the delay information has no impact on the perceived duration of the wait, it typically has an impact on both the acceptability of the wait and the affective response to waiting (Hui and Tse [48]). Moreover, the announcements are usually helpful because they provide customers with a sense of progress during their waiting experiences (Munichor and Rafaeli [65]).

1.2 Focus, Aim, and Organization

Focus. In this survey, we restrict attention to papers where the firm decides on whether and how to communicate delay information to its customers. In particular, customers cannot search for this information, nor can they acquire

it themselves, e.g., as in Hassin and Haviv [37], Hassin and Roet-Green [41], and Yang et al. [91]. We restrict attention to sharing waiting-time information, and do not include papers which consider alternative forms of shared information, e.g., on the service quality or the service rate; e.g., see Hassin [35] and Veeraraghavan and Debo [81]. Also, because the queueing-theoretic literature which studies properties of waiting times is vast, we restrict attention to papers that relate specifically to delay announcements. The first mathematical model of a queueing system with rational customers is Naor [67], where the queue is assumed to be observable to customers; the first unobservable model is studied in Edelson and Hilderbrand [25]. Numerous extensions to both models have been considered in the queueing-games literature, and the majority of those papers are relevant, albeit indirectly, to the problem of sharing delay information in queueing systems; see Hassin and Haviv [39] and Hassin [36] for comprehensive surveys. Of those papers, we only consider ones which compare, in a broad sense, the observable and unobservable models. Essentially, this amounts to quantifying the value of sharing delay information.

Aim and organization. In this survey paper, our objectives are: (i) To classify and systematically review the relevant papers; (ii) to identify the main challenges entailed in the different approaches to the problem; (iii) to synthesize some key findings of the literature; and (iv) to identify gaps in the literature, and formulate research directions which would be interesting to investigate in the future. In order to achieve our goals, we first need to organize the surveyed literature along key axes. Here, we classify papers based on the assumptions that they make about the way customers respond to the announcements. In particular, we identify three main literature sub-streams:

- In §2, we survey papers which model customers as *queued entities* that do not react to the announcements received. This literature stream focuses, for the most part, on investigating the accuracy of various wait-time predictors in alternative queueing models.
- In §3, we survey papers which model customers as utility-maximizing, forward-looking, *decision makers*. The focus of this stream is, primarily, on designing the effective control of the system, by studying the timing, breadth, and granularity of the shared information.
- In §4, we survey papers which assume that customers respond to the announcements, but where the specifics of the customer decision-making process, leading up to that response, are not modelled. In other words, customer response is assumed to be *exogenous*. The main objective of this stream is to study the existence and properties of resulting system equilibria, where announcements and actual delays "are close" in some sense.

For each one of the above categories, we precede our review of the relevant papers by outlining the main challenges in the corresponding stream of the literature. We conclude this survey by discussing some key ideas that hold broadly (§6), and listing some potential future research directions (§7).

2 Customers as Queued Entities

We begin by surveying papers which treat customers as queued entities that do not react to the announcements that they receive. For the most part, this branch of the literature focuses on studying ways of accurately predicting future waiting times. This is important for two main reasons: (i) from a practical perspective, systematically making inaccurate announcements may lead to customer distrust in those announcements and, ultimately, customer dissatisfaction with the service provided; and (ii) from an analytical perspective, studying waiting times in queueing systems allows for the derivation of structural results which are useful for our general understanding of those models.

In broad terms, two types of methods are typically used for predicting waiting times: Queueing-theoretic and data-based. For queueing-theoretic methods (§2.2), the focus is on systematically considering alternative queueing models, and studying the accuracy of various real-time delay predictors in those models. The predictors may exploit different types of information about the state of the system at the time of the announcement, e.g., the queue length or the history of recent delays. Relying on data-based methods for delay prediction (§2.3) is relatively recent, and it usually allows for superior predictive power.

2.1 Snapshot of the Main Challenges

To measure accuracy, one must first decide on an appropriate measure. Typically, average measures of accuracy are used, e.g., the mean-squared error (MSE), which incorporates both the variance of the estimator and its bias. Under the MSE criterion, the conditional expectation of the waiting time, given some state information, is the most accurate prediction (there is no bias in this case). However, calculating such expected values is generally hard, and there is usually a need to resort to alternative predictions. The relative MSE, which is equal to the MSE divided by the expected waiting time, is useful for a relative measure of accuracy. One can also rely on accuracy measures which penalize overestimation and underestimation, e.g., by using a newsvendor-like objective where different costs are assigned to each.

Assessing the predictive power of alternative estimators is usually done through a combination of analytical and numerical methods. On one hand, deriving closed-form expressions for prediction errors allows for an understanding of the dependence of those errors on alternative model parameters; on the other hand, detailed simulation studies allow for the extension of theoretical results to realistic settings which are not amenable to direct analysis.

To illustrate the complexity in doing direct analysis, let us consider an announcement which is equal to the delay of the Last customer to have Entered Service (LES) at the time of arrival of the new delayed customer. In what follows, we deliberately keep our exposition at a high level to convey key intuition. The LES announcement is accurate if the (stochastic) state of the system that the LES customer encounters upon arrival, e.g., the queue length,

is "not too different" from the state that the new customer, to whom the announcement is made, encounters. In other words, we need to determine whether the time scale at which the state of the system changes is "much larger" than the magnitude of the LES delay; if so, then the state of the system would not change considerably during the LES delay, i.e., the LES delay should be an accurate prediction. Because doing direct analysis is prohibitively difficult, there is a need to resort to approximations. This is often possible by relying on a many-server heavy-traffic framework, where results on asymptotic accuracy are derived. There is no single way of defining asymptotic accuracy; usually, a properly scaled sequence of differences between wait-time estimators and corresponding delays is shown to converge to 0, e.g., in a distributional sense. Importantly, one must first decide on an appropriate asymptotic regime.

To describe large systems, which are usually of primary interest, one alternative is to consider the Quality-and-Efficiency-Driven (QED) or Halfin-Whitt regime (Halfin and Whitt [33]; Garnett et al. [28]) which strikes a balance between service quality and operational efficiency. To describe a system where waiting times are long, one can focus on the Efficiency-Driven (ED) regime instead (Whitt [87]). Analysis in the QED regime is simplified for two main reasons: (i) the system exhibits economies of scale so that, asymptotically, waiting times are negligible, and (ii) a snapshot principle (Reiman [70]) holds, under certain conditions, so that the state of the system during the waiting time of a delayed customer changes negligibly. When the system is overloaded, fluid-model approximations and ED diffusion-scale refinements perform remarkably well, and are typically used to establish asymptotic accuracy.

2.2 Queueing Methods for Delay Prediction

Because there is no universal "most accurate" predictor, i.e., one which performs well in all queueing contexts, Whitt [86] systematically explores alternative ways of predicting waiting times in a multi-server queueing model with multiple classes, under certain distributional assumptions, by exploiting various levels of system-state information. The types of information considered involve the queue length, individual customer abandonment and service rates, remaining service times of customers in service, etc. Full cumulative distribution functions of customer waiting times are estimated in each case, through either exact analysis or approximations.

Following up on Whitt [86], in a series of papers Ibrahim and Whitt [51, 52, 53, 54] investigate the asymptotic accuracy of alternative real-time delay announcements, based on either the queue length or the history of delays, in queueing systems with several realistic features, such as time-varying arrivals and general distributional assumptions. The predictors that Ibrahim and Whitt consider are all single-number estimates, e.g., the mean of the wait-time distribution conditional on the queue length seen, or LES. For the most part, they consider the MSE criterion for accuracy and rely on a many-server heavy-traffic framework to: (i) derive approximations for MSE-minimizing con-

ditional expected wait-time values, given system-state information, which serve as new announcements, and (ii) quantify the accuracy of the various announcements considered. They substantiate their theoretical results with an extensive simulation study, and formulate general insights on the usefulness and limitation of each type of delay prediction.

Ibrahim and Whitt focus solely on single-class systems. The performance of LES in multi-class systems is considered numerically in Thiongane et al. [80]: The authors use simulation to explore the accuracy of the LES predictor in the context of a Markovian multi-server, multi-class, system with abandonment. They explore the accuracy of LES-based announcements, including the weighted average of LES predictions, as well as predictors exploiting both the queue length and the LES delay. Bassamboo and Ibrahim [12] study the performance of LES with multiple classes as well, and provide theoretical support to some of the numerical observations in Thiongane et al. [80]. Nakibly [66] also considers a multi-class context, and allows for heterogeneous, class-dependent, service rates. She considers both exact and approximate methods. For example, in a two-server queueing system with a non-preemptive priority discipline and exponential class-dependent service times, she describes the waiting-time distribution using difference equations and a matrix geometric method. She also considers an iterative algorithm to approximate that distribution in more complex models with multiple priorities and many servers.

For an alternative measure of accuracy, Jouini et al. [56] consider a newsvendor problem cost function instead, which allows to penalize overestimation and underestimation of delays using different cost parameters. They consider a multi-class queue with a priority service discipline and time-varying arrival rates. They empirically validate their theoretical results using data from a network of real-life call centers. In such a network, determining the number of servers available at every time epoch is difficult to do. In a system with both time-variations and an unknown number of servers, they propose simple approximations for wait-time moments. They consider approximating the corresponding wait-time distributions by using Erlang and Normal distributions with those matched moments, and find optimal announcements from these distributions. Finally, they take their results to data: They quantify the performance of their predictors, along with a benchmark mean-delay predictor. They find that Erlang-based predictions have a superior performance.

2.3 Data-Based Methods for Delay Prediction

There has been recent interest in using data-mining techniques for delay prediction in service systems. There are several papers which focus solely on data-mining methods for wait-time prediction, e.g., in healthcare settings (Pereira et al. [69]), or transportation systems (Demir and Demir [18]). In contrast, we focus here on papers which emphasize the importance of combining both queueing-theoretic and data-mining methods.

Senderovich et al. [72, 73, 74] introduce a novel framework which combines process-mining techniques, machine-learning algorithms, and queueing-theoretic results to predict waiting times in service queues. Single-class systems are considered in Senderovich et al. [73], and multi-class systems in Senderovich et al. [74]. The authors consider various predictors, including delay-history-based predictors, such as LES. Such predictors are termed "snapshot" predictors because their asymptotic accuracy in certain queueing contexts is substantiated by Reiman's snapshot principle (Reiman [70]). They also consider two average predictors, one which averages over the entire history of delays, and another which clusters waits according to k loads, using $k-$means clustering. In general, snapshot predictors are found to be accurate in single-class settings, consistently outperforming average predictors. In a multi-class setting, snapshot predictors and regression-based methods yield good performances. Senderovich et al. [73, 74] are based on the analysis of call-center data. Senderovich et al. [72] focuses on a healthcare setting instead. In particular, the authors rely on predictors which combine patient information, e.g., previous visits and other related information, with real-time congestion measures, such as the current number of patients and recent lengths of stay. The proposed prediction method is shown to have superior performance.

Ang et al. [7] also consider a healthcare setting, and use data sets from four hospitals. They, too, emphasize a message similar to Senderovich et al. [72, 73, 74]: Combining queueing-theoretic results with data-mining techniques leads to superior predictive performance. They introduce a novel estimation method, Q-Lasso, which is inspired by both queueing theory and the Lasso method of statistical learning. In particular, they consider a queue-length-based predictor, which is equal to the ratio of the queue length to the processing rate, as a covariate in the Q-Lasso method. The authors find that the Q-Lasso method consistently outperforms other prediction methods such as rolling-average methods. The authors also implement their method in a hospital, and discuss related implementation challenges.

3 Customers as Decision Makers

In this section, we survey papers which treat customers as forward-looking, utility-maximising, decision makers. The announcements do not directly impact payoffs in the system i.e, a customer is not compensated for following or discarding the wait-time information. Moreover, the objectives of the different players, i.e., the firm and its customers, are usually not perfectly aligned. For example, the firm may want to increase throughput, whereas customers may want reduced waiting times. Hence, the main premise in that literature is that the impact of the announcements arises indirectly in equilibrium. There are many related questions: Does an equilibrium exist? If so, is it unique? Assuming that a unique equilibrium exists, what are the values of different objectives (revenue, social welfare, customer utility, etc.) at equilibrium? Most importantly, what is the impact of sharing delay information?

3.1 Preliminaries: The Classical Framework

The classical queueing system is an $M/M/1$ model. The first-come-first-served (FCFS) discipline is considered, and there is unlimited waiting space. Arrivals are according to a Poisson arrival process with rate λ, service times are independent and identically distributed (i.i.d.) with rate μ, and there is a single server. Customers are delay sensitive, and we let C denote the waiting cost per time unit for a customer (which is assumed to be paid when the customer enters service). Customers also receive a reward R from service.

In Naor [67], a customer inspects, upon arrival, the queue length (number of customers in the system) and decides whether to join or balk. An individual joins a queue of size i if, and only if, her expected utility $R - \frac{C(i+1)}{\mu} \geq 0$. The equilibrium joining strategy, i.e., individual optimizing strategy, is a threshold-based strategy where customers who observe n customers in queue upon arrival join if, and only if, $n + 1 \leq n_e$, where $n_e \equiv \lfloor R\mu/C \rfloor$. The social benefit, per unit of time, assuming a threshold joining strategy with threshold n is given by $\lambda(1 - p_n)R - Cq$, where p_n is the stationary probability of finding n in the system, given a maximum queue length of n, and q is the expected queue length. A pure threshold socially-optimal strategy exists, and Naor [67] shows that the social benefit attains its maximum at a value $n_s \leq n_e$. Rooted in this classical result, a general theme in the queueing-games literature is that the selfish behavior of utility-maximizing customers leads to sub-optimal equilibrium solutions compared to the socially-optimal solution. The aim is then to investigate ways of restoring the imbalance. In Naor's framework, by imposing an appropriate admission fee, i.e., a static, queue-length independent, price, θ, customers can be motivated to adopt the threshold n_s instead of n_e. The toll may also be set from a revenue maximizer's objective, i.e., to maximize $\lambda(1 - p_n)\theta$. In this case, the fee levied by the manager is too high, i.e., $n_r \leq n_s \leq n_e$, where n_r is the corresponding equilibrium threshold.

Edelson and Hilderbrand [25] consider the basic unobservable model, where customers do not observe the queue length upon arrival, and make joining decisions based on the expected waiting time. Customers may either join the queue, not join, or adopt a mixed strategy where they join with probability q. It is found that a unique equilibrium strategy exists, and that it is based on the value of R: if R is "low", then no customer joins; if R is intermediate, then customers adopt a mixed strategy with joining probability $q_e \equiv \mu C/R$; and, if R is large, then everyone joins. The social benefit function attains its maximum at a value q_{soc} such that $q_{soc} \leq q_e$. Thus, as in the observable case, individual optimization leads to queues that are longer than socially desired, but the gap can be corrected by imposing an appropriate admission fee. We note that the objectives of a profit maximizer and the social planner coincide.

3.2 Snapshot of the Main Challenges

With endogenous customer response, studying implications on different objectives is not easy. A first-order issue is to decide on an appropriate objective. For example, an engineer may care about throughput, whereas an economist may care about social welfare. Moreover, different objectives may be affected by the delay information in similar ways, but not always. For example, a naive view may assume that an increase in throughput, i.e., number of served customers, must correspond to an increase in waiting times. However, this need not be the case. Indeed, real-time delay information usually allows for a better matching between supply and demand, so that we may concurrently have increased throughput and shorter waiting times. Moreover, such results tend to be intimately tied to the specific modelling assumptions made.

To illustrate the complexity in this line of research, we consider the basic question: How does revealing information about the queue length, i.e., providing delay information, affect throughput and social welfare? For throughput: In the observable case, we know that revealing information would incite customers to join when the queue length is short, and deter them from joining when the queue length is long. In the unobservable case, where customers make their joining decisions based on the expected waiting time, they would join more if the system is, overall, not highly congested. Now, let us compare the observable and unobservable cases: It is not clear what the aggregate effect on throughput should be. Revealing the queue length may induce more customers to join, but if the system is, overall, lightly congested, then it may also deflect some customers who encounter an "exceptionally" long queue. The reverse argument holds when the system is heavily congested. Thus, it seems that no general statement can be made, and that the load in the system should play a role. For social welfare: We know from both Naor [67] and Edelson and Hilderbrand [25] that customers create negative externalities on other customers, and that they may join both observable and unobservable queues when it is not socially optimal for them to do so. Thus, it is not clear what the aggregate impact of revealing queue-length information on social welfare would be. In general, more complex issues should be considered, such as the granularity of the delay information (going beyond the reveal/ do not reveal dichotomy above), as well as the timing and breadth of the shared information. The literature that we survey next addresses such issues.

3.3 To Reveal or Not to Reveal? Observable versus Unobservable Queues

We begin by surveying papers which compare the observable and unobservable systems, i.e., address whether or not to reveal queue-length information.

Social welfare, revenue maximization, and throughput. Hassin [34] studied the impact of information suppression from both the social planner's and revenue maximizer's perspectives. In both cases, two quantities play a central role: (i)

the potential arrival rate, λ, and (ii) the value of service, relative to the cost of waiting, $\nu_s \equiv R\mu/C$. Hassin [34] compares profits, under profit-maximizing admission fees, in the observable and unobservable cases. He finds that if customers are "very" sensitive to delay, $\nu_s \leq 2$ i.e., $C \geq R\mu/2$, then it is optimal to reveal the queue length for all $\lambda > 0$. However, if customers are not very delay sensitive, $\nu_s > 2$, then there exists a threshold, Λ_R, such that it is only optimal to reveal the queue length for $\lambda > \Lambda_R$. The intuition behind these results is as follows: When λ is large, many customers would opt to balk based on average wait-time information, which is high because λ is high. In this case, disclosing the queue-length information encourages more customers to join in low-congestion states. While it is true that it also discourages customers from joining highly-congested states, the key is that these customers would have balked anyway in the unobservable case; thus, revealing information helps the firm. We now turn to the social welfare results. First, we note that the problem would be straightforward if a social welfare maximizing fee can be imposed. In this case, revealing delay information can only help the social planner since, in the observable case, a customer would enter only when it is socially desirable to do so, but this is not the case in the unobservable model. The more challenging case is when pricing cannot be socially controlled, e.g., because price regulation is not desirable, but information suppression can be socially controlled. Under the assumption of a revenue-maximizing toll, the values of ν_s and λ play similar roles, but the threshold on λ, Λ_S, is different and it is shown that $\Lambda_S < \Lambda_R$. Thus, a social planner may want to reveal the queue length when it is not optimal for a revenue maximizer to do so, i.e., for $\Lambda_S < \lambda < \Lambda_R$. However, it is never optimal to suppress information when a revenue maximizer voluntarily chooses to reveal it, i.e., for $\lambda > \Lambda_R$.

Chen and Frank [17] study how information suppression impacts throughput. Intuitions similar to the ones in Hassin [34] continue to apply, so we will be brief. In particular, for a fixed admission fee, the role played by the system's load is prominent. On one hand, if the arrival rate is low, in particular $\lambda < \Lambda^*$, then customers may be turned away by real-time queue-length information while they would have joined with a (low) average wait-time information. This implies that $\lambda_O < \lambda_U$, i.e., the effective joining rate is smaller in the observable system than in the unobservable system. On the other hand, if the arrival rate is high, in particular $\lambda > \Lambda^*$, then $\lambda_O > \lambda_U$. Shone et al. [76] take a different view and focus on the situation where the decision of a service provider to reveal the queue-length information *does not* affect throughput. Shone et al. [76] assume out the possibility of optimizing the admission fee. They compare the observable and unobservable systems in terms of joining rates, both individually optimal (selfish) and socially optimal (altruistic), in addition to various other system performance measures. The authors derive necessary and sufficient conditions for the equality of equilibrium selfish and altruistic joining rates between the observable and unobservable systems, and show that both equalities cannot simultaneously hold. Shone et al. [76] also observe that the decision of whether or not to reveal the queue length depends strongly on ν_s, as was observed in Chen and Frank [17].

A network of providers. The papers above focus on a setting with a single provider. Singh et al. [77] consider a competitive environment with two service providers instead. These providers may choose to diffuse different levels of information, either real-time or historical. The paper studies the first mover's benefit, i.e., the first provider to announce real-time information. It considers two parallel $M/M/1$ queues, in a multi-period setting, where one provider announces the real-time queue length, and the other provider announces the expected delay of the previous period. For a performance comparison, the authors consider the market share and the expected delay, and customers join the lower-delay alternative. The authors find that the benefit of being the first mover depends on the service capacity. In particular, for the lower-capacity provider, being the initiator in announcing real-time information increases the market share and reduces delays. However, the same does not hold for the higher-capacity provider, where results are mixed. The authors also find that social welfare always increases when there is benefit on market share and delay.

Dong et al. [20] also consider a network setting with multiple providers, but they focus on a network of hospitals instead. In particular, they study, in the context of an empirical investigation, the impact of delay announcements on coordination in the network. Coordination is measured through the correlations of delays between hospitals: There is synchronization if those correlations are positive. This observation is rooted in a queueing-theoretic result which establishes that the Join-the-Shortest-Queue (JSQ) discipline synchonizes queues in the system. Indeed, if customers check the delay information, then it is reasonable to assume that they would join the shortest queue, which would then lead to synchronization. Thus, exploring the impact of delay information reduces to studying correlations between the waiting times at adjacent hospitals. By relying on data of real-life announcements and patient response (measured through online searches), the authors investigate whether the announcements do indeed impact the behavior of patients. They provide empirical evidence that this is indeed the case. They also conduct an extensive numerical study to investigate how sensitivity of customers to delays, the load of the system, and the heterogeneity between hospitals, impact the synchronization level in the system. They show that using average-wait predictors may lead to oscillations in the system, where customers systematically flock to one of the two queues; this numerical observation is studied in Pender et al. [68].

3.4 Granularity, Timing, and Breadth of the Delay Information

The papers above consider either full revelation or no revelation of real-time system-state information. However, there are other considerations, such as the timing, granularity, and breadth of the shared delay information. We now survey papers which study decisions pertaining to those characteristics.

"Discrete" information: High and low announcements. The idea that full information may not be necessary, and that a discrete high-low type of announce-

ment may suffice, already follows immediately from Naor [67]. Indeed, in the observable case, customers follow a threshold-type joining decision; this indicates that only the information on whether or not the queue length exceeds a threshold, L, should suffice. Because this information structure is much simpler, there is interest in studying it. We note that setting $L = 0$ corresponds to the unobservable model in Edelson and Hilderbrand [25].

Altman and Jimenez [6] consider high-low announcements when there is no pricing decision. First, the authors consider that the value of L is fixed (not necessarily at optimum). In the social planner problem, they optimize the probabilities of accepting an arrival if the queue length is below or above L. Next, they consider the individual optimization problem where utility-maximizing customers make their joining decisions, and investigate the ensuing equilibrium. In both problems, the optimal admission strategy has the form of either accepting all arrivals when the queue length is below L, or rejecting all arrivals when it is above L. The authors also show that imposing a socially-optimal L value in the individual optimization problem does not lead to the socially optimal outcome. Hassin and Koshman [40] consider a similar setting as in Altman and Jimenez [6], albeit with pricing decisions. In particular, customers are charged p_L when the queue length is below L, and p_H otherwise. Hassin and Koshman [40] demonstrate how to obtain the maximum value of social welfare in Naor's model by using their coarse dynamic pricing scheme.

The above two-signal strategy arises at equilibrium in Allon et al. [5]. In this paper, the authors relax two fundamental assumptions: (i) that the firm is truth-telling in revealing information, and (ii) that the information shared is quantifiable and verifiable by customers. As such, they allow for a richer information set which also includes intentional vagueness: A firm is intentionally vague when it provides the same announcement in different states of the system. They show that even though the information provided to customers is nonverifiable, it can improve the profits of the firm and the expected utility of customers. The incentives of the firm and its customers are neither perfectly misaligned (they both prefer shorter waits), nor perfectly aligned (the firm benefits from higher throughput, whereas the customers do not). This misalignment between the firm and its customers plays a key role in the analysis: Depending on its level, different equilibria emerge. Of particular interest are equilibria with influential cheap talk, i.e., ones where the firm can induce distinct customer actions based on different unverifiable messages.

The authors show that any pure-strategy equilibrium with influential cheap talk can be mapped into a single threshold queue-length level, i.e., the firm would provide two signals to indicate whether the level of congestion is above or below that level. Because it suffices to restrict attention to threshold-based equilibria, the misalignment between the firm and its customers is defined as follows. In an observable system, customers behave as in Naor [67] and have a threshold-based joining strategy, q^*. If the firm had full control over customers, it would impose a different joining threshold, $\hat{q}$, instead. The misalignment, ϕ, is defined as $\phi \equiv q^* - \hat{q}$. If $\phi = 0$ (perfect alignment) or $\phi < 0$ (customers are impatient), then an influential pure-strategy equilibrium exists. This pure

strategy can then be mapped into a two-signal equilibrium, as above. However, if $\phi > 0$, then such an equilibrium cannot exist because customers would always deviate from the firm's threshold for balking. The authors also show that there always exists a babbling equilibrium, where the firm provides meaningless signals and customers ignore them, and a most informative equilibrium, where the firm may use intentional vagueness to incite customers to join states that they would not join otherwise.

Different levels of information. We now turn to the literature investigating the problem of finding the "best" type of delay information to share. Duenyas and Hopp [21] investigate that problem in a manufacturing setting. Each customer who places an order generates a reward for the firm, and there is a penalty for being late (per unit time exceeding the quoted lead time). In response to a quoted lead time, a, each customer places an order with probability $p(a)$. Duenyas and Hopp [21] derive an optimal quote which maximizes the expected profit (revenue minus penalty cost), under both infinite ($G/G/\infty$) and finite ($G/G/1$) capacity settings. In the infinite-capacity case, the optimal quote does not depend on the current backlog in the system. In the finite-capacity alternative, the optimal lead-time quoting policy is state-dependent and increasing in the state, i.e., the higher the congestion, the higher the lead-time quote. Specifically, a profit-maximizing firm should give granular, state-dependent, information rather than rely on a coarse information-sharing scheme.

In their model, Duenyas and Hopp [21] trade the reliability of the quoted delay for maximizing throughput: While there is a penalty for being late, the firm is not, otherwise, restricted in the quote that it provides, i.e., it is not constrained to being reliable. In contrast, Dobson and Pinker [19] consider a similar problem but assume that the firm must provide reliable quotes: The state-dependent lead-time quote provided, l_i, depends on the number i of customers in the system, and is a fractile from the conditional wait-time distribution which must be met $(100\tau)\%$ of the time. In other words, letting W_i denote the conditional steady-state waiting time, we must have that $\mathbb{P}(W_i \leq l_i) = \tau$. The proportion of customers who join the system, in response to l_i, is given by $\alpha(l_i, \tau)$. Dobson and Pinker [19] compare alternative scenarios, S_k, which reflect different levels of information granularity: For scenario S_k, customers are provided with a state-dependent announcement l_i for $i < k$, and with a static announcement for $i \geq k$. Increasing k amounts to increasing the granularity of the delay information. The authors derive a sufficient condition under which sharing more information increases throughput, and emphasize that this need not always be the case. Importantly, they demonstrate that higher throughput may also be associated with lower expected waiting times, and less variable waits, because the delay information deters customers from joining highly-congested states, and encourages customers to join low-congestion states. They also highlight the importance of customer heterogeneity, i.e., the extent to which different information granularity leads to different demand rates: The more the heterogeneity, the higher the throughput, i.e., the higher the value that can be derived from quoting lead times.

The role played by customer heterogeneity is also central in the work of Guo and Zipkin. Guo and Zipkin [29] consider three levels of information: (1) no information, (2) partial information i.e., queue length upon arrival, and (3) full information i.e., exact waiting time. For performance measures, they consider throughput and the expected customer utility. Customers are assumed to be heterogeneous in their delay costs. Specifically, each arriving customer has a cost type, θ, which is drawn from a continuous and bounded distribution, H, and density function, h. There is also a basic cost function, $c(w)$, associated with a wait w. Thus, the cost incurred by a θ-customer who is delayed for w is equal to $\theta c(w)$. Different levels of delay information incite more or less customers to join. The information provided also segments customers depending on their delay sensitivity: A customer who joins under one type of information, may balk under another type. Guo and Zipkin [29] demonstrate that both system throughput and customer utility, under different information levels, are impacted by the shape of the customer-delay distribution. Depending on that distribution, they characterize conditions under which information helps either the customers or the service provider. In particular, if the density, h, does not decrease too sharply, then throughput increases with more information. Intuitively, this is so because if h decreases too sharply, then too few customers join in high-congestion states. Also, if h does not rise too sharply, then average customer utility increases with more information. Intuitively, this is so because if h rises too sharply, then too many customers join in low congested states. The main takeaway is that more information may or may not be beneficial, depending on the distribution of customer delay sensitivity. In subsequent papers, the above results are generalized to systems with phase-type service times (Guo and Zipkin [30]), different levels of information (Guo and Zipkin [31]), and alternative cost functions (Guo and Zipkin [32]).

In a series of papers, Burnetas and Economou [15], Economou and Kanta [23, 24], and Economou et al. [22], the authors quantify the impact of state information on system dynamics under various assumptions. Burnetas and Economou [15] consider an $M/M/1$ queue with setup times. In particular, when a new customer arrives to an empty system, the server requires an exponentially-distributed time with rate θ before beginning service. At time t, the state of the system is described by the pair $(N(t), I(t))$ where $N(t)$ is the number of customers in the system and $I(t) = 0$ or 1 is the state of server (idle or busy, respectively). Customers may be exposed to different levels of information about the system, corresponding to four cases: (i) fully observable, where customers observe both $N(t)$ and $I(t)$; (ii) almost observable, where customers observe only $N(t)$; (iii) almost unobservable, where customers observe only $I(t)$; (iv) fully unobservable, where customers do not observe either $I(t)$ or $N(t)$. In all cases, customer equilibrium strategies are analyzed, as well as the stationary behavior in the system and the social benefit for all customers. Through a series of numerical experiments, the authors show that whether or not the social welfare in the system is benefitted by the additional information depends on both θ and λ. However, in general, the difference in social benefit is small between the fully and almost observable cases, but there

may be significant differences between the observable and unobservable systems. In other words, the information about the server state yields marginal benefit compared with the queue-length information. Economou et al. [22] consider an extension of Burnetas and Economou [15] where both general service and general setup times are allowed. Economou and Kanta [24] assume that the waiting space is divided into compartments, to be served sequentially in increasing order, and joining customers may know either the compartment number (but not their position in the compartment that they join) or their position within a compartment (but not the compartment number). Both information levels correspond to partial information since customers do not fully observe the system state in either case. For a frame of reference, if a customer knows both the compartment index and the compartment position, then the model reduces to the model in Naor [67], whereas if neither are known then the model reduces to the model in Edelson and Hilderbrand [25]. Economou and Kanta [23] and Wang and Zhang [82] assume that the server may break down and require repair. The time to repair is considered to be equal to 0 in the former, and is exponentially distributed in the latter. The authors in those two papers compare two levels of information: (i) fully observable, where customers know both the queue length and the state of the server, and (ii) partially observable, where customers know only the queue length. Both papers compare equilibrium threshold balking strategies in their contexts.

Timing and breadth. The question of when to make a delay announcement, and the extent to which information should be shared, have also been investigated in the literature. He and Down [43] rely on both heavy-traffic analysis and simulation to study performance in a queueing system where only a fraction of customers are informed about waiting times. Specifically, they consider two customer classes and two server pools. Dedicated customers in each class can only be served by one of the two pools, e.g., because of a language requirement. A fraction of customers is flexible, and may choose one of the two server pools depending on which has the shortest queue. He and Down [43] focus on the expected waiting time for both classes, and demonstrate that "a little flexibility goes a long way" in that delay information (the queue length) significantly improves performance even when a small proportion of customers are informed about waiting times. They also address the question of information updating by considering, numerically, a setting where the mean waiting time is updated periodically, and customers use the most recent update in making their joining decisions. They show that there could be significant degradation in performance if the delay information is not updated frequently enough, and the system may experience oscillation behavior because customers herd together for one queue for a period of time.

Hu et al. [45] also address the question of the breadth of the information shared. They consider a setting where only a fraction of customers are informed about the queue length in the system. Informed customers make their joining decisions based on the observed queue length. Uninformed customers make their joining decisions based on the expected waiting time in the system.

The fraction of informed customers is assumed to be exogenous. Informed customers join the system in accordance with the threshold joining policy in an observable queue, as in Naor [67]. Uninformed customers randomize their joining decisions. Uninformed customers indirectly influence informed customers by influencing the distribution of the queue length in the system. The authors find that, in systems which are not under very low loads, informing a fraction of customers about real-time delay information increases either the throughput or the social welfare. Their results depend on both the offered load in the system and the joining behavior of uniformed customers. To relate their results to Chen and Frank [17]: They find that when the offered load is low enough, throughput decreases with the information. Similarly, if the offered load is high enough, then throughput increases with the information. However, in the intermediate region for the offered load, throughput is maximal if only a fraction of customers are informed. Also, while the standard view, as in Hassin [34], is that social welfare is always improved by revealing the queue, the authors demonstrate that when the offered load is high enough, it is optimal to have only a fraction of informed customers, i.e., social welfare does not always increase by revealing the queue length to everyone. In short, the presence of uninformed customers improves throughput under low offered loads, and increases social welfare under high offered loads.

Despite its practical importance, the question of timing of the announcements remains understudied, with the vast majority of papers assuming that the announcement is given immediately upon arrival of the delayed customer. At a high level, the tradeoff is as follows: Postponing the announcement allows the firm to make a more informed decision about whether or not to admit the customer. With more information at its disposal because of the delay in making the announcement, the firm should benefit. However, postponing the announcement also means potentially keeping customers longer in queue. Thus, it is not clear whether a firm would want to resort to this postponement. Allon and Bassamboo [4] address this question in the context of an unobservable $M/M/N$ queue; the model specifics are, otherwise, similar to Allon et al. [5]. The authors focus on identifying conditions under which influential cheap talk emerges in equilibrium. To model the system with postponed announcements, they consider a two-stage system. The first stage, which models e.g., a call center's IVR, is an infinite-server queue which is essentially a delay station. The second stage is an $M/M/N$ queue: Upon entry to this $M/M/N$ queue, the firm makes a non-verifiable cheap-talk type of delay announcement. The authors characterize the optimal admission policy for the firm in the second stage, and demonstrate that it is of a threshold type where the threshold depends on the number of customers in the first stage. They also characterize the set of possible equilibria in the delayed cheap talk game, and compare these to the non-delayed game. They show that such a comparison is complex: The firm may or may not benefit, i.e., create credibility and impact customer behavior, from delaying the delay information.

Pender et al. [68] also consider the impact of delaying the delay announcements. Specifically, they study the oscillation behavior observed in both He

and Down [43] and Dong et al. [20]. They use two deterministic fluid models to examine the effect of providing customers with delayed delay information. In particular, they consider two systems: System I consists of two infinite-server queues where arriving customers receive delayed information about the queue length. The delay in information is quantified by a deterministic parameter Δ. Customers choose which queue to join depending on the delayed delay information that they receive, in accordance with a multinomial logit customer choice model. By analyzing the dynamics of the resulting fluid model, the authors demonstrate that there is asynchronous behavior between the two queues if Δ is large enough, i.e., there are systematic oscillations and no stable equilibrium. System II also consists of two infinite-server queues, but the delay information is in the form of a time-average of the queue-length information in a window of length Δ instead. In this case as well, the authors demonstrate a similar asynchronous behavior between the two queues if the window over which the average is taken is long enough.

Roet-Green and Hassin [71] also consider a setting where customers learn delayed information about the queue length in the system but, contrary to Pender et al. [68], the delay in information is assumed to be random (exponentially distributed), corresponding to the travel time needed for a customer to join the queue after the delay information is received. In other words, customer joining decisions are not instantaneous. A customer joining strategy is a vector that assigns a probability of traveling to each possible queue length. Because the travel time is not negligible, a customer who had decided to join a system based on "old" queue-length information, may decide to balk upon arrival to the system if the real-time queue length is too long. Thus, customer decisions are made at two successive epochs. The authors investigate the structure of a symmetric Nash equilibrium. They find that customers often adopt a double-threshold strategy: customers travel when the queue length is short, balk or mix between balking and traveling when the queue length is at an intermediate length, and travel when the queue length is long. The intuition is that a customer who observes an intermediate queue assumes that previous customers must have observed short queues, and are now on their way. Thus, the system's congestion is likely to soon increase and, consequently, the customer decides to balk. The intuition is reversed when a customer observes a long queue: In this case, that customer assumes that previous customers must have observed an intermediate queue and balked. Thus, the congestion in the system is likely to soon decrease, and the customer decides to join the queue. The authors also demonstrate that social welfare may be higher under the no-information model than under the delayed information model.

Hu and Wang [46] consider a setting where customers share queue-length information with each other. Because information is shared at the arrival epoch of an arriving customer, it constitutes lagged information for a future customer who wishes to join the system based on this "historical" information. Customers decide to join or balk based on previous information, but do not update their decisions upon arrival to the system because they do not observe the queue length in the second stage, unlike in Roet-Green and Hassin [71].

Indeed, they observe the queue length only upon entering the system. The authors investigate how this shared information structure affects throughput, expected queue length, and social welfare in the system, and draw comparisons between the full-information and no-information models. They find that (i) throughput under shared information is less than that under full information; (ii) the expected queue length under shared information is less than that under full information; and (iii) social welfare may be lower or higher under shared information, depending on the offered load in the system.

3.5 Joint Optimization: Announcements and Other Controls

Because delay announcements are levers of control in the system, it is natural to investigate how a manager may jointly optimize the announcements with other levers of control, such as staffing and scheduling decisions.

Armony and Maglaras [8, 9] study joint routing and delay-announcement decisions in the context of a call center which offers a call-back option to delayed customers. Specifically, callers are informed, upon arrival, of their predicted waiting time for real-time service, and a delay guarantee for postponed service. There is a continuum of delay-sensitive customer types, and customers assign utilities to joining either queue, and join the queue corresponding to the highest utility. The problem is how to provide accurate delay estimates and decide on an accompanying routing rule which guarantees that the postponed service is offered within the specified deadline. This problem is analytically difficult to solve, primarily because future arrivals from the postponed service may affect the waiting times of customers who are already in queue. Thus, the authors focus on the many-server heavy-traffic Halfin-Whitt regime instead. Under this regime, the authors show that using a local version of Little's law, i.e., announcing the queue length encountered upon arrival divided by the arrival rate, is asymptotically consistent (it becomes accurate in large systems) under a threshold-type routing rule which is asymptotically compliant (satisfying the delay guarantee constraint). Specifically, the manager gives priority to real-time service customers, so long as the queue-length for the postponed service does not exceed a give threshold. While Armony and Maglaras [9] focuses on steady-state delay information, Armony and Maglaras [8] considers state-dependent delay information instead. In comparing the performance of the system with steady-state or state-dependent delay information, the authors show that state-dependent information increases resource utilization while improving the quality of service for real-time service.

Yu et al. [94] also consider a setting where a profit-maximizing firm uses the announcements in conjunction with optimizing a routing rule, but where customer types are unobservable to the manager. Because customers are heterogeneous in both their delay costs and the values drawn from service, the firm may gain from customer segmentation through a priority service discipline. There is information asymmetry in the model: While the firm has private information about the congestion level in the system, customers have

private information about their types. Since information on customer types is not observable by the firm, the announcements play a dual role: They inform customers about upcoming (expected) delays, and they are means to eliciting information about customer types. In other words, the priority discipline used by the firm depends on the announcements given. The authors examine the ability of the firm to sustain an equilibrium with influential cheap talk in the above setting, and distinguish between two cases, depending on whether the two customer classes considered have homogeneous or heterogeneous holding costs. In the homogeneous case, they show that the firm can achieve its unconstrained first-best profit, where it has both full information and full control over customers, through the provision of delay announcements. In particular, a partial segmentation of the customer population may be sufficient to achieve maximal profit. Moreover, under certain conditions, not differentiating customers at all may be the profit-maximizing strategy. In the heterogeneous case, the firm can no longer achieve its first best trough the announcements. Nevertheless, it can improve its profits by giving priority to customers who receive the highest announcements. The authors also characterize babbling equilibria in the system, where no credible information is shared with customers so that the state of the system and the announcement given by the firm are independent; they also compare babbling equilibria to influential equilibria where the firm communicates credible information to customers. They find that providing credible delay information always increases the firm's profit, but may improve or hurt the expected total customer utility.

Ibrahim [49] also takes the view that the announcements can be used as a control tool which can be optimized jointly with other controls. In particular, the focus there is on a queueing system where the number of servers is random. This setting arises in sharing-economy applications, e.g., because of the self-scheduling behavior of work-from-home call-center agents. Because agents show up at random, there are congested periods in the system. Because of this congestion, the abandonment distribution plays an important role. In particular, it can be controlled, via delay announcements, to alleviate the cost of self-scheduling. The author studies how to control the announcements, along with other tools, namely the compensation offered to agents and the staffing level in the system, in order to minimize costs.

3.6 Empirical Studies

The literature above is analytical in nature. The recent availability of granular data, e.g., at the call-by-call level in call centers, has made it possible to study changes in customer behavior, in response to the announcements. We now recap the main results from those papers.

Early empirical evidence which illustrates how customers update their patience times in response to delay announcements, in call centers, can be found in Mandelbaum and Zeltyn [63] and Feigin [26]. Akşin et al. [3] undertake a more detailed empirical study to explore the impact on customer behavior

and, in turn, on system performance, due to the announcements. The authors begin by providing empirical evidence, using a Cox regression analysis, substantiating the impact of the announcements on the abandonment behavior of (call center) customers. Their data set has two priorities, and the announcements are equal to the queue position or the elapsed waiting time of the longest waiting customer; they are also made sequentially over time. The study reveals that both the composition and sequence of the announcements has an impact on customer abandonment behavior, and that customers who receive longer announcements, or see a deteriorating delay condition (increasing announcements during their wait), abandon earlier. The impact of the announcements is also affected by the priority class of the customer.

In order to explore the operational impact of the announcements, the authors use a structural estimation approach: They model callers' abandonment decisions as in the optimal stopping time model introduced in Akşin et al. [2]. Specifically, time is divided into periods, and a customer makes a decision on whether or not to abandon at the beginning of each period. Customers are heterogeneous in both the rewards that they receive from service and their per-unit waiting costs (both of these are drawn from lognormal distributions). The announcements received impact the abandonment distribution of callers which, in turn, impacts their decisions on staying or reneging, sequentially over time. The parameters of that endogenous model for caller abandonment are estimated from data, for each priority class. In order to study the impact of the announcements, the authors assume a setting where customers receive only one announcement upon arrival. By relying on the approximation in Whitt [88], they characterize the equilibrium that arises in the system in steady state, where the equilibrium is defined as one where the distribution of waiting times based on the optimal stopping time model coincides with the distribution of the waiting time using the approximation from Whitt [88]. Through a simulation study, Akşin et al. [3] then study the operational impact of the announcements. Their main conclusions are as follows: (i) delay information helps customers make better decisions in the sense that callers who receive a long (short) delay announcement abandon more and faster (less and slower); (ii) the impact of the announcements is strongest when the state of the system is congested; and (iii) the increased granularity of the wait-time announcement (exact queue length position versus range for the number in queue) leads to a smoother change in caller behavior.

Yu et al. [92] also adopt an empirical approach in studying the impact of delay announcements on customer patience. They begin by introducing the concepts of informative and influential announcements. An informative announcement is one that carries information about the current congestion level in the system, i.e., one where longer delays do indeed correspond to larger announcements. An influential announcement is one where the patience of customers changes in response to the announcements. By statistically comparing the survival distributions of customers, the authors find that the impact of the announcements is ambiguous: Some announcements are influential and/or informative, whereas others are not. This prompted the authors to under-

take a deeper investigation into the dynamics of the performance impact of the announcements; they did so by relying on a structural estimation approach.

The structural model is as follows: Customers may return multiple times and, at each return, receive multiple delay announcements during their wait. At each announcement epoch, the caller revisits their decision of staying until service or reneging. Customers are heterogeneous, but their heterogeneity is modelled through their cost-reward ratio rather than separately through their service rewards and waiting costs. The cost-reward ratios and variance of idiosyncratic shocks are then estimated from data. The authors consider two models: (i) a base model where customers update their beliefs about offered waits using the announcements received; (ii) a refined model where not only customer beliefs but also the waiting costs of customers are impacted by the announcements. The authors find that their second model explains the ambiguous impact on customer impatience observed earlier in their data analysis. In particular, they show that while the cost-reward ratio decreases in the offered wait associated with the announcements ("I waited so long already, so why not wait a little longer?"), the variance of the idiosyncratic shocks increases. This dual effect explains the nontrivial impact of the announcements on customer behavior. The authors then explore, through a simulation study, what managerial implications can be drawn from their analysis. In particular, they find that providing delay announcements leads to an increase in the surplus of customers (surplus is equal to reward minus waiting cost), and that less refined delay information (in the form of three signals on the congestion of the system) may lead to higher customer surplus than more granular information.

Yu et al. [93] undertake a field experiment in an Israeli bank's call center to explore the loss aversion of customers, in time, and its dependence on the delay information available. Specifically, customers who receive delay announcements typically form a reference point based on the announcement received. If the actual waiting time experienced is smaller than that reference point, then the time difference is considered a gain. If the actual waiting time experienced is larger, then the time difference is considered a loss. Loss aversion means that customers value lost time more than they value gained time. Customers are either provided with accurate, inaccurate, or no announcements. By using a structural model to infer the customers' value of time (the abandonment behavior is modelled through an optimal stopping time problem), the authors find that customers indeed exhibit loss aversion, and that this is independent of the correctness of the delay information given. (Loss aversion is measured through an increase in the per-unit waiting cost after the announcement.) However, the accuracy of the delay announcement does have an impact on the reference point formed. Specifically, with accurate information, the reference point coincides with the delay information given, whereas with inaccurate information, customers use the observed average delay as a reference point instead. This contradicts the standard viewpoint that firms should give an inaccurate but high announcement to make the customers "feel better about their waits". Indeed, the analysis suggests that customers may disregard such inaccurate announcements but retain their loss aversion.

In a related paper, Webb et al. [83] rely on a proportional hazards model for the hazard rate of the abandonment distribution instead. The covariates used in that model include the gain and loss in time effects due to the announcements. In particular, the announcement creates a reference point which is the expectation of the wait-time for service. The authors find that a model in which customers react to the announced value of the first announcement, and in which reference points are induced by the first two announcements, is the best fit to their data. They also find that customers are loss averse, that they fall for sink cost effects, and that a higher announcement leads to more abandonment. Finally, they study implications on staffing decisions, and find that firms who take behavioral implications of the announcements into account can significantly reduce their staffing levels.

4 Exogenous Changes in Customer Behavior

In this section, we survey papers which model customer response to the announcements, but do not explicitly model the dynamics of the customer decision process leading up to that response. In other words, changes in customer behavior are assumed to be exogenous. While the literature in §3 focuses on modelling customer joining/balking decisions only, the papers in this section consider changes in customer abandonment behavior as well and focus, for the most part, on the accuracy of the announcements in light of those changes.

4.1 Snapshot of the Main Challenges

Studying the accuracy of delay announcements, when customers respond to these announcements, is challenging. Indeed, changes in customer impatience affect system dynamics and, in turn, the future announcements made. For example, if customers abandon faster because of high announcements, then future waiting times, and future announcements which depend on those waiting times, should be shorter. Thus, studying the accuracy of the announcements involves characterizing an equilibrium in the system. At a high level, an equilibrium must correspond to the long-run performance in the system, where the average announced delay coincides with the average experienced delay.

First, it is not clear whether such an equilibrium exists, or if it is unique; indeed, there may be multiple equilibria and the system may exhibit oscillations between those equilibria. Second, even when a unique equilibrium exists, it is not clear how to specify that the announcement and the corresponding delay, which are both random variables, coincide in that equilibrium, e.g., this could be in expectation, in distribution, or asymptotically when scaled in an appropriate way. Third, it is not clear how stochastic fluctuations around the equilibrium affect the system's performance and the accuracy of the announcements. Even under Markovian assumptions, explicit analysis of the underlying birth-and-death process is analytically complex. This is so because the

transition rates of the birth-and-death chain would all be dependent on the announcements. Therefore, analysis is typically done in an asymptotic heavy-traffic regime instead. However, establishing asymptotic accuracy is not easy primarily because it may be that the underlying stochastic processes, e.g., the queue-length process, do not even converge. Even if the underlying processes do converge, then the analysis is complicated by the state-dependent nature of the arrival and abandonment rates in the system, due to the announcements.

4.2 Accuracy and Performance Impact

One fundamental idea is that the announcements help by deterring the most impatient customers from waiting, i.e., by converting late abandonment into early balking. Indeed, since those customers would have abandoned anyway, inciting them to abandon immediately upon arrival, i.e., balk, should help in reducing congestion in the system while not affecting throughput.

Replacing exponential reneging with balking. We begin with the case where customers who receive delay information consider it to be truthful, know their personal preferences, and are able to decide, upon arrival, whether they would be willing to wait at all. In this case, all reneging is replaced by balking because of the announcement. Whitt [85] adopts this view, and compares two single-class $M/M/s/r$ queueing systems (where r denotes the maximum queue length allowed) with reneging and balking. In particular, Model 1 assumes that customers balk with a given probability and otherwise join the system and may renege after some time. Model 2 assumes that customers are given system-state information upon arrival, e.g., the queue length. In Model 2, all reneging is replaced with balking at arrival. Because of the dynamics of customer response, and conditional on the system state seen upon arrival, a customer does not take other customers' actions into account when making her own decision to join or balk. By analyzing general birth-and-death processes, with announcement-dependent rates, Whitt [85] shows that the number of customers in Model 1 is larger in Model 2 in the likelihood-ratio stochastic ordering sense. In other words, the announcements lead to an improvement in performance by converting reneging after some delay with balking upon arrival. Jouini et al. [57] extend Whitt [85] by considering a system with two customer classes and a non-preemptive priority service discipline. In a model where customers replace subsequent reneging with balking upon arrival, as in Whitt [85], the authors do analysis to derive balking probabilities and moment expressions for the virtual waiting times of the high and low priority customers.

In practice, delay announcements do not convert reneging entirely into balking (Mandelbaum and Zeltyn [63], Feigin [26]). Indeed, it seems more common that the most impatient customers balk in response to the information, while more patient customers elect to stay but update their patience levels, depending on the announcement. If the announcement is long, then there will be more balking, and less subsequent reneging, and vice versa. Thus, there is

a tradeoff between reneging and balking, based on the announcement. This is one of the main ideas in Jouini et al. [55]. The authors consider a delay announcement which is equal to a fixed percentile of the waiting time, conditional on the queue length seen upon arrival, and study how varying that percentile, or coverage β, impacts performance in the system. Jouini et al. [55] consider three models: Model 0 assumes that the delay information is exact, and arriving customers respond to that information by balking upon arrival if their patience falls below that threshold; there is no subsequent balking in the system. This model is in the same spirit as Whitt [85], and it is argued that this is indeed a reasonable model when customers fully trust the information that they are given. Model 1 assumes no delay announcement, and that a higher proportion of customers balks upon arrival because of the lack of information; customers may later renege if their patience expires before reaching service. Model 2 introduces the idea of a coverage-based announcement, where the firm announces a given percentile of the waiting-time distribution. In this model, customers update their patience based on the announcements that they receive: The updated patience rate, γ', is equal to a combination of their individual patience before the announcement, and the delay information received (later approximated by an exponential distribution for the analysis). Under an exponential assumption on the announcement-dependent abandonment, the authors rely on the analysis of birth-and-death models to analyze the performance impact of the announcements. For consistency, the announcement given must coincide with the fractile of the stationary delay distribution. Thus, an equilibrium analysis is needed, and the announcement-dependent abandonment rate is derived based on a fixed-point algorithm. This algorithm reveals the dependence of γ' on β. Thus, varying β leads to different performance in the system. The authors find that, all else held constant, an announcement with more coverage leads to higher balking in lieu of late abandonment from the queue. However, through investigating the value of the "optimal" coverage (minimizing the balking probability, subject to a constraint on the reneging probability) an important insight is reached: More coverage, which is equivalent to more precise delay information, at the expense of a larger announcement, is not always better for the service provider. Indeed, this would depend on a host of factors, including the way in which customers react to the specific announcements that they receive.

Non-exponential but smooth abandonment. Armony et al. [10] go beyond the exponential assumption on the abandonment distribution, in response to the announcement. Direct analysis is hard, and the authors rely on two approximation methods to study the resulting equilibrium in the system: (i) a deterministic fluid model and (ii) an iterative numerical algorithm, based on Whitt [88], where general abandonment is approximated by Markovian abandonment with state-dependent rates. The authors focus on the performance impact of making the LES delay announcement. By analyzing the fluid model, they derive conditions on customer response to guarantee the existence and uniqueness of that equilibrium. In the fluid model, LES coincides at equilib-

rium with a fixed delay announcement (FD), equal to the average equilibrium delay. This motivates the authors to also consider an FD announcement, and they use simulations to study the equilibrium behavior with both LES and FD announcements in the $M/GI/n+GI$ model. They validate both approximation methods, and illustrate that the LES announcement is usually more effective, leading to smaller variance. Using the framework in Armony et al. [10], on can quantify the value of communicating delay information, e.g, by comparing the equilibrium that arises with performance in a system without announcements. This performance impact depends on the assumptions made on the way customers respond to the announcements.

Armony et al. [10] do not discuss the accuracy of the individual announcements, which involves quantifying the stochastic fluctuations around equilibrium. This is done, in a similar setting, in Ibrahim et al. [50]. The authors demonstrate that the LES announcement, with customer response to the announcements, is asymptotically accurate in both the quality-and-efficiency driven and efficiency-driven regimes. A main technical issue in the analysis is demonstrating that the stochastic fluctuations around the equilibrium in the system (when it exists and is unique) would not drive the system out of that equilibrium, thus guaranteeing accuracy.

Abandonment "jumps": Going beyond the fluid model. Armony et al. [10] illustrate that the fluid model may not be accurate when the abandonment response to the announcements is not smooth, e.g., when there is an announcement-dependent "jump" in abandonment, which is consistent with empirical evidence. To analyze the system with such jumps necessitates going beyond the fluid approximation, i.e., a more refined approximation is needed. Such an approximation is presented in Huang et al. [47]. Because the announcements play a role in altering customer abandonment, it is conceivable that jointly optimizing the control of announcements along with the staffing level would lead to staffing levels that are different than in the absence of announcements. Huang et al. [47] are the first to show this by considering an overloaded $GI/M/s+GI$ queue where they jointly optimize the staffing level and the timing of the announcements, subject to quality-of-service constraints. The announcement-dependent hazard rate of the abandonment distribution is assumed to be discontinuous. In particular, they consider two types of delay announcements, corresponding to two types of responses. The first is similar to Armony et al. [10], where customers who hear an announcement upon arrival have a changed abandonment response at the point of the announcement, as well as balking upon arrival in response to the announcement. The second type of announcement is made during the waiting time, leading to an abrupt increase in the likelihood of abandonment at the announcement epoch. The objective of the paper is to quantify the impact of the non-smooth change in abandonment on system performance and operational decisions. To do so, the authors introduce an approximation based on scaling the patience-time distribution. They substantiate the accuracy of their refined approximation, demonstrate that there is an $\mathcal{O}(\sqrt{\lambda})$ reduction in the staffing level due to the announcements, and

show that the optimal timing of the announcement coincides with the fluid offered waiting time.

5 Alternative Contexts

Throughout this paper, we focused mainly on delay announcements in the context of customer service systems. In this section, we include a brief discussion of related problems in various different contexts.

Production systems: Spearman and Zhang [78] consider two problems in single-class production systems with multiple stages and unlimited buffer capacities. Problem I investigates the optimal lead-time policy when the objective is to minimize the average due-date minus the arrival date subject to a constraint on the fraction of tardy jobs. Problem II uses the same objective but imposes a constraint on the average job tardiness instead. Spearman and Zhang [78] show that the solution to the first problem results in "unethical" practices (quoting an unattainable lead time), whereas the solution to problem II is more appropriate. More recent references on jointly optimizing the lead-time quote along with pricing include Çelik and Maglaras [16] and Wu et al. [89].

Retail: Lu et al. [60] conduct an empirical study to investigate the effect of waiting times on customer purchasing behavior in a retail store. They find that customer purchasing behavior tends to be influenced by the length of the queue rather than by the processing speed (i.e., the waiting time itself). Moreover, they find that customer sensitivity to waiting is heterogeneous.

Kidney transplants: Bandi et al. [11] study how to optimally allocate kidneys to transplant patients. They consider a multiclass, multiserver queuing system, where heterogeneous patients await kidneys of different qualities. They use a robust optimization solution methodology and develop an online tool to deliver fine-grained predictions on how long to wait for an appropriate kidney to arrive, which incorporate information about the patient's kidney-quality preferences. They calibrate their model using detailed historical data, and numerically illustrate the superiority of their prediction method.

Ticket queues: Xu et al. [90] investigate the management of ticket queues where customers do not observe the queue length but can draw a ticket upon arrival. Customers balk if the difference between the ticket number and the displayed number exceed a customer's patience. The authors develop tools to investigate performance in the system, based on a Markov chain model, and compare physical and ticket queues: They show that the balking probabilities can be very difference between the two settings.

Transportation: Gal et al. [27] consider the problem of predicting travel time in transportation systems. They propose a prediction engine which combines both queueing-theoretic and machine learning methods to predict the travel time of buses between a source and a destination. To predict bus travel times, the authors consider first "snapshot-based" predictions where the travel times of recent buses are considered. They also consider regression techniques to predict bus travel times, and integrate the snapshot predictor into the regression model, in the same spirit as Ang et al. [7]. They find that prediction methods which combine both the snapshot principle and regression tree techniques outperform the separate snapshot predictors or regression tree methods.

6 Discussion

In this section, we identify some key concepts that can be synthesized based on the surveyed literature. The list below is not meant to be exhaustive; rather, its purpose is to formulate some broad insights based on that literature.

Heterogeneity can be exploited through the announcements. In a setting where delay information is shared with customers, one general insight is that alternative levels of heterogeneity can be effectively managed, through the provision of delay announcements, to lead to superior outcomes. In that sense, the announcements may be viewed as a type of pricing tool which segments the customer population in an appropriate way, e.g., as in priority pricing (Adiri and Yechiali [1]). For one example, with a homogeneous customer population, the manager can benefit from "creating" heterogeneity by controlling the breadth of shared real-time congestion information; indeed, having both informed and uninformed customers can lead to improved throughput, social welfare, or operational performance (Hu et al. [45]). For another example, heterogeneity in customers' tolerances for waiting can be effectively managed through the provision of delay announcements to lead to increased throughput and social welfare (Dobson and Pinker [19], Guo and Zipkin [29]). For yet another example, unobservable heterogeneity in customer types (reward from service and waiting cost) can be managed by the announcements to lead to increased profits (Yu et al. [94]). For a last example, the heterogeneity in service capacities, between two competing service providers, makes sharing real-time delay information beneficial, for both market share and operational performance, for the low-capacity firm (Singh et al. [77]).

More information is not always better. One may have different objectives in mind when assessing the value of providing delay information, and those objectives may be impacted by that information in different ways. While the value of information provision is usually context-dependent, one general principle is that providing more information need not always lead to improved performance, and may even be detrimental. From a human psychology angle, customers do not always prefer more granular information (Hui and Tse [48],

Ang et al. [7]). For both social welfare and throughput, less granular delay information may be beneficial (Hassin [34], Chen and Frank [17], Burnetas and Economou [15], Guo and Zipkin [29, 30], Roet-Green and Hassin [71], Hu and Wang [46], etc.). Moreover, nonverifiable and non-quantifiable information may improve both the firm's profit and the expected utility of customers (Allon et al. [5]). Finally, under certain conditions, providing delay information may make the system more volatile, and can lead to longer delays on average (He and Down [43], Armony et al. [10], Dong et al. [20], Pender et al. [68]).

Of course, providing delay announcements helps in many cases. In particular, another general insight is that providing real-time delay information usually yields the greatest benefit, e.g., for profit, social welfare, and throughput, when the system experiences heavy congestion (Chen and Frank [17], Hassin [34], He and Down [43], Hu et al. [45], etc.). Also, from an accuracy perspective, various delay predictors can be proved to have a superior performance under such high-congestion conditions as well, particularly when the system is large (Ibrahim and Whitt [51, 52]).

There is no single "best" announcement. There is no universal best way to predict waiting times, and the accuracy of a specific announcement depends on both the amount of state information available, and the specific modelling context (Whitt [86]). Thus, there is a need to consider several such contexts and to study performance under each specific setting. There are also different measures of performance, ranging from the average error, e.g, using the MSE, to penalising under or overestimation (Jouini et al. [56]).

In broad terms, under the MSE criterion, and conditional on some system-state information, e.g., the queue length, the conditional expectation of the waiting time, given that information, is the most accurate prediction. While calculating conditional expectations is possible under certain conditions, it is, generally, a difficult task. Moreover, those resulting conditional expected values tend to perform poorly when the specific modelling assumptions under which they were derived fail to hold (Ibrahim and Whitt [51]). Thus, one needs to consider alternative, and simpler, ways to predict delays, e.g., by exploiting the recent history of delays in the system (Armony et al. [10]). Such delay-history-based predictions can perform remarkably well, e.g., in large heavily-congested systems with or without customer abandonment, even when customers respond to the announcements (Ibrahim et al. [50]). There is also some empirical evidence substantiating their good performance in practice (Senderovich et al. [73, 74]). However, they do not perform well in other settings, such as when the system is small or lightly loaded (Yu et al. [93], Thiongane et al. [80]), or under time-varying conditions (Ibrahim and Whitt [53]). The main takeaway is this: While the literature does not give us a conclusive answer as to what type of announcement to use under all circumstances, it does provide valuable insights on the appropriateness of various announcements in different settings.

Data methods and queueing-theoretic methods are complementary. The recent proliferation of empirical studies, in the context of delay announcements,

prompts one to evaluate the alternative methods that are used to address that problem. In broad terms, the literature ranges from analytical work, typically substantiated by simulation-based results (Whitt [85], Armony and Maglaras [8], etc.), to empirical work in the context of a well-defined structural model (Akşin et al. [3], Yu et al. [92], Webb et al. [83], etc.), to work which relies, for the most part, on data-mining methods (Ang et al. [7], Senderovich et al. [73, 74, 72]). Each body of work is important in its own right, and it is crucial to emphasize the complementarity of those different approaches. Indeed, while relying on queueing models is instrumental to gain insight into performance and, importantly, allows for a mathematical framework through which controlling that performance is made possible, queueing-theoretic methods typically lack robustness in that they remain intimately tied to the specific technical assumptions under which the analysis is derived.

Empirical studies in the context of a well-defined structural model on customer utility have been instrumental in both validating existing models on customer response to the announcements, and extending those models as well. Grounded in both empirical evidence and theoretical analysis, they enable a better management of delay announcements in practice.

Data-mining methods are clearly superior in terms of accuracy. Thus, if accuracy is the sole objective in mind, then there seems to be little value in going beyond them. However, data-mining techniques are limited in that they are "black-box" techniques that do not, in general, further our understanding about the dynamics of the system. Recently, the combination of those two frameworks (queueing and data-based) has been advocated in several papers (Ang et al. [7], Senderovich et al. [73, 74, 72]). Indeed, the delay predictors in those papers are inspired by both queueing-theoretic methods and data-mining techniques, and are shown to yield superior performance with real-life data sets. In the same spirit, Bassamboo and Ibrahim [12] propose a framework to quantify the accuracy of delay announcements across different queueing models. That framework enables an easier assessment of that accuracy with real-life data, which circumvents the need to fit entire queueing models to data in order to gain insight into performance.

7 Future Research Directions

In this section, we identify some "macro-level" themes that we believe would be interesting to investigate in future research.

Bridging the psychological and the operational. As mentioned in Bitran et al. [14], there is a general need to narrow the gap between mathematical models of customer response in service systems, and the complex reality of human behavior and psychology. The body of literature devoted to analyzing customer response to the announcements generally assumes that changes in customer behavior arise from individual customers maximizing their expected utilities

from service and waiting. Some papers have challenged whether such an approach is always appropriate. For example, Guo and Zipkin [29] indicate that relying on utility-based approaches may lead to counter-intuitive results, such as customers preferring more congested states. In the same spirit, Allon et al. [5] indicate that customers may not be expected-utility maximizers and may, e.g., prefer accuracy over no accuracy, or information over no information. To wit, extant experimental work from the psychology and marketing literatures offers important insight on how customers perceive and react to both having to wait for service, and to receiving delay announcements while waiting (§1.1). There remains ample opportunity to design more sophisticated models which incorporate such psychological features, to test the validity of those models with data, and to study implications on decision-making in the system.

One recent work in that vein is Yu et al. [93], which tests the loss aversion of customers (to waiting) by conducting a field experiment in the context of a call center. Webb et al. [83] study implications on operational decision-making with similar behavioral features. Another relevant work, though not specifically related to delay announcements, is Yuan et al. [95]. In this paper, service providers share a common entertainment option, which alleviates the cost of waiting on their customers, but compete on other service dimensions. This duality between cooperation and competition is termed co-opetition. The authors demonstrate that a service provider's profit can increase when engaging in co-opetition. In other words, Yuan et al. [95] quantify how a psychological dimension, i.e., making the customer waiting experience more pleasant, can indeed influence traditional operational measures, such as the firm's profit. Further studies, in the same spirit, are interesting venues for future research.

Alternative designs for delay announcements. In the literature on delay announcements, it is commonly assumed that a single delay announcement is given to customers, that the manager decides on whether or not to provide the announcement, and that the delay information is given immediately upon arrival. Recent work has begun challenging those assumptions (Allon et al. [5], Pender et al. [68], Roet-Green and Hassin [71], Hu and Wang [46]), primarily on issues concerning the timing of the announcement, and how the announcements are actually diffused to the customer population, e.g., whether this is done by the manager or by the customers themselves.

Recent technological advances have made it possible for firms to obtain a wealth of data about individual customers, and to track the evolution of the service experience, in real time. This opens up an opportunity for a better segmentation of (heterogeneous) customers, e.g., via targeted delay announcements, and a study of the implications of such segmentation on performance in the system. Optimizing the granularity of the information shared with such heterogeneous customers, potentially sequentially during their stay, is an interesting topic for future research. Indeed, experimental evidence suggests that people value a sense of progress during their waiting times, which can be made possible through the announcements; e.g., see Munichor and Rafaeli [65]. Moreover, by targeting customers with (multiple) different announce-

ments, the firm can incite different abandonment behaviors. While models for rational customer abandonment have been advanced in some papers (Hassin and Haviv [38], Mandelbaum and Shimkin [62], Haviv and Ritov [42], Shimkin and Mandelbaum [75], etc.), and have been substantiated empirically in others (Akşin et al. [2], Yu et al. [92]), systems with endogenous abandonment, which is dependent on delay announcements, remain understudied in general. In a context with announcement-dependent abandonment, jointly optimizing the provision of announcements and the scheduling of those impatient customers (e.g., in the spirit of Bassamboo and Randhawa [13]) would be possible. Further exploration of, e.g., the design of a system with multiple announcements, the study of the dynamic impact of such sequential announcements on customer behavior, and the analysis of corresponding implications on the operational management of the system, and on various related objectives, remain interesting venues for future research.

Towards building a service science. In this survey paper, we reviewed papers taking different approaches to the effective management of delay announcements in service systems. The overall objective of that rich body of work is to build a service science. With that goal in mind, it is important to systematically study different queueing models with various complexities, and to paint a complete picture of the impact and accuracy of delay announcements. Therefore, it is important to mention that several model extensions remain underexplored, despite their prevalence in practice. Here is a non-comprehensive list of such extensions: non-stationary and non-Poisson arrival models; alternative service disciplines (beyond FCFS); multiple classes with heterogeneous service rates and/ or heterogeneous abandonment rates; queueing networks; queues where capacity is uncertain e.g., due to the self-scheduling behavior of agents; queues where different levels of information are missing, e.g., on service, arrival, and abandonment rates, and on customer types and classification, etc.

References

1. Adiri, Igal, Ury Yechiali. 1974. Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research* **22**(5) 1051–1066.
2. Akşin, Zeynep, Barış Ata, Seyed Morteza Emadi, Che-Lin Su. 2013. Structural estimation of callers' delay sensitivity in call centers. *Management Science* **59**(12) 2727–2746.
3. Akşin, Zeynep, Barış Ata, Seyed Morteza Emadi, Che-Lin Su. 2016. Impact of delay announcements in call centers: An empirical approach. *Operations Research* **65**(1) 242–265.
4. Allon, Gad, Achal Bassamboo. 2011. The impact of delaying the delay announcements. *Operations research* **59**(5) 1198–1210.

5. Allon, Gad, Achal Bassamboo, Itai Gurvich. 2011. we will be right with you: Managing customer expectations with vague promises and cheap talk. *Operations research* **59**(6) 1382–1394.
6. Altman, Eitan, Tania Jimenez. 2013. Admission control to an m/m/1 queue with partial information. *International Conference on Analytical and Stochastic Modeling Techniques and Applications*. Springer, 12–21.
7. Ang, Erjie, Sara Kwasnick, Mohsen Bayati, Erica Plambeck, Michael Aratow. 2015. Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management* **18**(1) 141–156.
8. Armony, Mor, Constantinos Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**(4) 527–545.
9. Armony, Mor, Constantinos Maglaras. 2004. On customer contact centers with a call-back option: customer decisions, routing rules, and system design. *Operations Research* **52**(2) 271–292.
10. Armony, Mor, Nahum Shimkin, Ward Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81.
11. Bandi, Chaithanya, Nikolaos Trichakis, Phebe Vayanos. 2017. Robust wait time estimation in resource allocation systems with an application to kidney allocation. Northwestern University, working paper.
12. Bassamboo, Achal, Rouba Ibrahim. 2017. Delay announcements in service systems: When is the average wait good enough? Northwestern University, working paper.
13. Bassamboo, Achal, Ramandeep Singh Randhawa. 2016. Scheduling homogeneous impatient customers. *Management Science* **62**(7) 2129–2147.
14. Bitran, Gabriel, Juan-Carlos Ferrer, Paulo Rocha e Oliveira. 2008. Om forummanaging customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing & Service Operations Management* **10**(1) 61–83.
15. Burnetas, Apostolos, Antonis Economou. 2007. Equilibrium customer strategies in a single server markovian queue with setup times. *Queueing Systems* **56**(3) 213–228.
16. Çelik, Sabri, Costis Maglaras. 2008. Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Science* **54**(6) 1132–1146.
17. Chen, Hong, Murray Frank. 2004. Monopoly pricing when customers queue. *IIE Transactions* **36**(6) 569–581.
18. Demir, Engin, Vahap Burhan Demir. 2017. Predicting flight delays with artificial neural networks: Case study of an airport. *Signal Processing and Communications Applications Conference (SIU), 2017 25th*. IEEE, 1–4.
19. Dobson, Gregory, Edieal Pinker. 2006. The value of sharing lead time information. *IIE Transactions* **38**(3) 171–183.
20. Dong, Jing, Elad Yom-Tov, Galit B Yom-Tov. 2017. The impact of delay announcements on hospital network coordination and waiting times. Northwestern University, working paper.

21. Duenyas, Izak, Wallace Hopp. 1995. Quoting customer lead times. *Management Science* **41**(1) 43–57.
22. Economou, Antonis, Antonio Gómez-Corral, Spyridoula Kanta. 2011. Optimal balking strategies in single-server queues with general service and vacation times. *Performance Evaluation* **68**(10) 967–982.
23. Economou, Antonis, Spyridoula Kanta. 2008. Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs. *Operations Research Letters* **36**(6) 696–699.
24. Economou, Antonis, Spyridoula Kanta. 2008. Optimal balking strategies and pricing for the single server markovian queue with compartmented waiting space. *Queueing Systems* **59**(3) 237–269.
25. Edelson, Noel, David Hilderbrand. 1975. Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society* 81–92.
26. Feigin, Paul. 2006. Analysis of customer patience in a bank call center .
27. Gal, Avigdor, Avishai Mandelbaum, François Schnitzler, Arik Senderovich, Matthias Weidlich. 2015. Traveling time prediction in scheduled transportation with journey segments. *Information Systems* .
28. Garnett, Ofer, Avishai Mandelbaum, Martin Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) 208–227.
29. Guo, Pengfei, Paul Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970.
30. Guo, Pengfei, Paul Zipkin. 2008. The effects of information on a queue with balking and phase-type service times. *Naval Research Logistics (NRL)* **55**(5) 406–411.
31. Guo, Pengfei, Paul Zipkin. 2009. The effects of the availability of waiting-time information on a balking queue. *European Journal of Operational Research* **198**(1) 199–209.
32. Guo, Pengfei, Paul Zipkin. 2009. The impacts of customersdelay-risk sensitivities on a queue with balking. *Probability in the engineering and informational sciences* **23**(03) 409–432.
33. Halfin, Shlomo, Ward Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations research* **29**(3) 567–588.
34. Hassin, Refael. 1986. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society* 1185–1195.
35. Hassin, Refael. 2007. Information and uncertainty in a queuing system. *Probability in the Engineering and Informational Sciences* **21**(03) 361–380.
36. Hassin, Refael. 2016. *Rational Queueing*. CRC Press.
37. Hassin, Refael, Moshe Haviv. 1994. Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying. *Stochastic Models* **10**(2) 415–435.
38. Hassin, Refael, Moshe Haviv. 1995. Equilibrium strategies for queues with impatient customers. *Operations Research Letters* **17**(1) 41–45.
39. Hassin, Refael, Moshe Haviv. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems*, vol. 59. Springer Science & Business Media.

40. Hassin, Refael, Alexandra Koshman. 2017. Optimal control of a queue with high-low delay announcements: the significance of a queue. Tel Aviv University, working paper.
41. Hassin, Refael, Ricky Roet-Green. 2017. The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research* **65**(3) 804–820.
42. Haviv, Moshe, Ya'acov Ritov. 2001. Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing systems* **38**(4) 495–508.
43. He, Yu-Tong, Douglas G Down. 2009. On accommodating customer flexibility in service systems. *Infor* **47**(4) 289.
44. Hornik, Jacob. 1984. Subjective vs. objective time measures: A note on the perception of time in consumer behavior. *Journal of consumer research* **11**(1) 615–618.
45. Hu, Ming, Yang Li, Jianfu Wang. 2017. Efficient ignorance: Information heterogeneity in a queue. *Management Science* .
46. Hu, Ming, Jianfu Wang. 2017. Efficient inaccuracy: Information sharing in a queue. University of Toronto, working paper.
47. Huang, Junfei, Avishai Mandelbaum, Hanqin Zhang, Jiheng Zhang. 2017. Refined models for efficiency-driven queues with applications to delay announcements and staffing. *Operations Research* .
48. Hui, Michael K, David K Tse. 1996. What to tell consumers in waits of different lengths: An integrative model of service evaluation. *The Journal of Marketing* 81–90.
49. Ibrahim, Rouba. 2017. Managing queueing systems where capacity is random and customers are impatient. University College London, working paper.
50. Ibrahim, Rouba, Mor Armony, Achal Bassamboo. 2017. Does the past predict the future? the case of delay announcements in service systems. *Management Science* **63**(6) 1762 – 1780.
51. Ibrahim, Rouba, Ward Whitt. 2009. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management* **11**(3) 397–415.
52. Ibrahim, Rouba, Ward Whitt. 2009. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science* **55**(10) 1729–1742.
53. Ibrahim, Rouba, Ward Whitt. 2011. Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. *Production and Operations Management* **20**(5) 654–667.
54. Ibrahim, Rouba, Ward Whitt. 2011. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations research* **59**(5) 1106–1118.
55. Jouini, Oualid, Zeynep Akşin, Yves Dallery. 2011. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* **13**(4) 534–548.

56. Jouini, Oualid, Zeynep Akşin, Fikri Karaesmen, Salah Aguir, Yves Dallery. 2015. Call center delay announcement using a newsvendor-like performance criterion. *Production and Operations Management* **24**(4) 587–604.
57. Jouini, Oualid, Yves Dallery, Zeynep Akşin. 2009. Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics* **120**(2) 389–399.
58. Katz, Karen L, Blaire M Larson, Richard C Larson. 1991. Prescription for the waiting-in-line blues: Entertain, enlighten, and engage. *MIT Sloan Management Review* **32**(2) 44.
59. Leclerc, France, Bernd H Schmitt, Laurette Dube. 1995. Waiting time and decision making: Is time like money? *Journal of Consumer Research* **22**(1) 110–119.
60. Lu, Yina, Andrés Musalem, Marcelo Olivares, Ariel Schilkrut. 2013. Measuring the effect of queues on customer purchases. *Management Science* **59**(8) 1743–1763.
61. Maister, David. 1985. The psychology of waiting lines. *The Service Encounter*. 113–23.
62. Mandelbaum, Avishai, Nahum Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36**(1) 141–173.
63. Mandelbaum, Avishai, Sergey Zeltyn. 2013. Data-stories about (im) patient customers in tele-queues. *Queueing Systems* **75**(2-4) 115–146.
64. McGuire, Kelly, Sheryl Kimes, Michael Lynn, Madeline Pullman, Russell Lloyd. 2010. A framework for evaluating the customer wait experience. *Journal of Service Management* **21**(3) 269–290.
65. Munichor, Nira, Anat Rafaeli. 2007. Numbers or apologies? customer reactions to telephone waiting time fillers. *Journal of Applied Psychology* **92**(2) 511.
66. Nakibly, Efrat. 2002. Predicting waiting times in telephone service systems. Ph.D. thesis, Technion–Israel Institute of Technology.
67. Naor, Pinhas. 1969. The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society* 15–24.
68. Pender, Jamol, Richard Rand, Elizabeth Wesson. 2017. Managing information in queues: The impact of giving delayed information to customers Cornell University, working paper.
69. Pereira, Sonia, Filipe Portela, Manuel Santos, Jose Machado, Antonio Abelha. 2015. Predicting pre-triage waiting time in a maternity emergency room through data mining. *Smart Health*. Springer, 105–117.
70. Reiman, Martin I. 1982. The heavy traffic diffusion approximation for sojourn times in jackson networks. *Applied probability computer science: the interface*. Springer, 409–421.
71. Roet-Green, Ricky, Rafael Hassin. 2017. Information can reduce social welfare when customers decide whether or not to travel to a queue. University of Rochester, working paper.
72. Senderovich, Arik, Matthias Weidlich, Avigdor Gal. 2017. Feature learning for accurate time prediction in congested healthcare systems. Technion, working paper.

73. Senderovich, Arik, Matthias Weidlich, Avigdor Gal, Avishai Mandelbaum. 2014. Queue mining–predicting delays in service processes. *International Conference on Advanced Information Systems Engineering*. Springer, 42–57.
74. Senderovich, Arik, Matthias Weidlich, Avigdor Gal, Avishai Mandelbaum. 2015. Queue mining for delay prediction in multi-class service processes. *Information Systems* **53** 278–295.
75. Shimkin, Nahum, Avishai Mandelbaum. 2004. Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems* **47**(1) 117–146.
76. Shone, Rob, Vincent Knight, Janet Williams. 2013. Comparisons between observable and unobservable m/m/1 queues with respect to optimal customer behavior. *European Journal of Operational Research* **1**(227) 133–141.
77. Singh, Siddharth Prakash, Mohammad Delasay, Alan Scheller-Wolf. 2017. Evaluating the first-movers advantage in announcing real-time delay information Carnegie Mellon University, working paper.
78. Spearman, Mark, Rachel Zhang. 1999. Optimal lead time policies. *Management Science* **45**(2) 290–295.
79. Taylor, Shirley. 1994. Waiting for service: the relationship between delays and evaluations of service. *The journal of marketing* 56–69.
80. Thiongane, Mamadou, Wyean Chan, Pierre L'Ecuyer. 2016. New history-based delay predictors for service systems. *Winter Simulation Conference (WSC), 2016*. IEEE, 425–436.
81. Veeraraghavan, Senthil, Laurens Debo. 2009. Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management* **11**(4) 543–562.
82. Wang, Jinting, Feng Zhang. 2011. Equilibrium analysis of the observable queues with balking and delayed repairs. *Applied Mathematics and Computation* **218**(6) 2716–2729.
83. Webb, Eric M., Qiuping Yu, Kurt M. Bretthauer. 2017. Linking delay announcements, abandonment, and staffing: A behavioral perspective. Indiana University, working paper.
84. Weinberg, Bruce. 2000. Don't keep your internet customers waiting too long at the (virtual) front door. *Journal of Interactive Marketing* **14**(1) 30–39.
85. Whitt, Ward. 1999. Improving service by informing customers about anticipated delays. *Management science* **45**(2) 192–207.
86. Whitt, Ward. 1999. Predicting queueing delays. *Management Science* **45**(6) 870–888.
87. Whitt, Ward. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* **50**(10) 1449–1461.
88. Whitt, Ward. 2005. Engineering solution of a basic call-center model. *Management Science* **51**(2) 221–235.

89. Wu, Zhengping, Burak Kazaz, Scott Webster, Kum-Khiong Yang. 2012. Ordering, pricing, and lead-time quotation under lead-time and demand uncertainty. *Production and Operations Management* **21**(3) 576–589.
90. Xu, Susan H, Long Gao, Jihong Ou. 2007. Service performance analysis and improvement for a ticket queue with balking customers. *Management science* **53**(6) 971–990.
91. Yang, Luyi, Laurens Debo, Varun Gupta. 2016. Trading time in a congested environment. *Management Science* **63**(7) 2377 – 2395.
92. Yu, Qiuping, Gad Allon, Achal Bassamboo. 2017. How do delay announcements shape customer behavior? an empirical study. *Management Science* **63** 1–20.
93. Yu, Qiuping, Gad Allon, Achal Bassamboo. 2017. The reference effect of delay announcements: A field experiment. Indiana University, working paper.
94. Yu, Qiuping, Gad Allon, Achal Bassamboo, Seyed Iravani. 2017. Managing customer expectations and priorities in service systems. *Management Science* .
95. Yuan, Xuchuan, Tinglong Dai, Lucy Gongtao Chen, Srinagesh Gavirneni. 2017. Co-opetition in service clusters with waiting-area entertainment. Harbin Institute of Technology, working paper.