Strategic customer behavior in a queueing system with alternating information structure

Yiannis Dimitrakopoulos[1], Antonis Economou[1], and Stefanos Leonardos[2]

[1]National and Kapodistrian University of Athens, Department of Mathematics, Panepistemiopolis, Athens 15784, Greece
[2]Singapore University of Technology and Design, 8 Somapah Rd, 487372 Singapore
{dimgiannhs, aeconom}@math.uoa.gr; stefanos_leonardos@sutd.edu.sg

June 14, 2019

Abstract: Strategic customer behavior is strongly influenced by the level of information that is provided to customers. Hence, to optimize the design of queueing systems, many studies consider various versions of the same service model and compare them under different information structures. In particular, two extreme versions are usually considered and compared: the observable in which customers are informed about the number of customers in the system and the unobservable in which they are only informed about the system parameters, e.g., arrival and service rates. In the present work, we study a model that bridges these two versions. More concretely, we assume that the system alternates between observable and unobservable periods. We characterize and compute customer equilibrium joining/balking strategies and show that the present model unifies and extends existing approaches of both heterogeneously observable models and models with delayed observations. More importantly, our findings indicate that an alternating information structure implies in general higher equilibrium throughput and social welfare in comparison to both the observable and unobservable cases. We complement our results with numerical experiments and provide managerial insight on the optimal control of the system parameters.

Keywords: Queueing Games; Strategic Customers; Equilibrium Strategies; Alternating Information Structure

1 Introduction

The strategic customer behavior in queueing systems has received considerable attention since the pioneering paper of Naor [20], who studied the M/M/1 queue from an economic viewpoint. Naor assumed that the customers are active entities who decide whether to join or balk after observing the queue length. He also considered the problem of a social planner and a monopolist who take into account the customer strategic behavior when they aim to optimize the social welfare and the profit respectively. The study of this observable version of the M/M/1 queue was subsequently complemented by Edelson and Hildebrand [6] who considered the same problems for the unobservable version of the system. In their model, the arriving customers are not allowed to observe the number of customers in the system and make their decisions relying solely on the knowledge of its operational and economic parameters. Since then, the literature has grown considerably. Hassin and Haviv [12] and Hassin [11] survey the basic methodology and results till 2003 and from 2003 till 2016 respectively. Stidham [24], in his monograph on optimal design

of queueing systems, provides a comprehensive overview of the various fundamentals models in the area.

One recurrent theme in the strategic queueing literature is the study of the effect of the level of information that is provided to the customers on their strategic behavior. This is an important theoretical issue per se, but it is also important for a social planner or a monopolist that are interested in the optimal design of a given system. What is interesting is that the effect of the level of information is ambiguous. Using the framework of the M/M/1 queue to reduce the complexity of the problem, existing studies demonstrate that more information can benefit or hurt the customers and/or the service provider, depending on various parameters and structural assumptions of the underlying model, see e.g., Hassin [10], Chen and Frank [3], Guo and Zipkin [8].

The results in this literature show that neither the observable nor the unobservable versions of the M/M/1 queue are preferable for the whole range of the underlying operational and economic parameters. To gain further insight, a number of authors studied the M/M/1 queue with strategic customers under information structures that lie between the observable and unobservable versions. To the best of our knowledge, there are three main ideas that have appeared in the literature that bridge the observable and unobservable versions of the M/M/1 queue: partially observable models [5, 9, 23, 17, 13], heterogeneously observable models [4, 15] and models with delayed observations [2, 14, 21]. We provide a brief overview of these results in Subsection 1.2.

In the present paper, we develop an alternative model that bridges observable and unobservable models and unifies preexisting approaches. Specifically, we consider a queueing system of M/M/1 type with strategic customers, which alternates between observable and unobservable periods: customers that arrive during observable periods see the number of present customers before making their joining/balking decisions, whereas customers that arrive during unobservable periods are only informed about the system parameters, e.g., arrival and service rates, but not about the actual queue length.

1.1 Motivation, objectives and contribution

The model of the present study is motivated by a number of different situations that arise in practice depending on whether the alternation between observable and unobservable periods is intentional or unintentional. A first such situation occurs when the information-providing mechanism has to respect some periodicity of the mode of operation of the system. For example, a given system may have to follow the day-night alternation and some of its features should be shut down in the night. An unintentional alternation between observable and unobservable periods occurs when the information-providing mechanism is unreliable. In this case, the system is observable as long as the mechanism operates properly, but it becomes unobservable when it fails and is under repair. Finally, a third case occurs when the administrator of the system intentionally stops to provide queue-length information for economic or other reasons, but resumes the information provision later. This is reasonable under various scenarios, e.g., if the information-providing mechanism is costly.

To understand how the information structure affects strategic customer behavior and system performance, we study these situations under the unified framework of a First-Come-First-Served (FCFS) M/M/1 queue that alternates between observable and unobservable periods. Upon arrival, customers decide whether to join or balk given the information that they receive from the system. In case that they enter the queue, they may renege (abandon the queue) any time later. Some basic considerations about the optimal equilibrium strategies are fairly straightforward in this setting. Customers that arrive at observable periods use Naor's [20] threshold strategy n_e: they join if the queue is less than a particular length and balk otherwise. Moreover, due to the FCFS discipline and the exponentially distributed service times, customers that enter during observable periods do not have an incentive to renege at any subsequent time.

In contrast, customers that arrive during unobservable periods, will enter with some probability q that depends on their expected payoff. Such customers may have an incentive to renege at the time of the first change to observable mode after their arrival instants. Regarding reneging, they again follow a threshold strategy with tolerance n_s (in terms of queue length) that is at least as high as Naor's threshold.

Based on these considerations, the search for equilibrium strategies can be restricted in the class of strategies that are parametrized by two non-negative integer thresholds n_e and n_s, and a probability q. We refer to such strategies as *potentially equilibrium strategies* (PES) and denote them by (n_e, n_s, q)-(PES). Assuming that the population of customers adopts a PES, the state of the system can be described by a Continuous Time Markov Chain (CTMC). This enables the computation of the expected net benefit of a tagged customer that joins during an unobservable period in terms of the steady-state distribution of the CTMC and in turn, the equilibrium joining probability q_e. More specifically, using that the number of customers in the system is stochastically increasing in q and that their net benefit is strictly decreasing in q, we derive that such an equilibrium joining probability q_e always exists, is unique and can be characterised via the system parameters. This concludes the technical analysis of the model and provides the required tractability for numerical experiments on its economic, operational and managerial aspects.

In this respect, our theoretical findings and experimental comparative statics indicate that the alternating information structure crucially affects system performance. To study the response of the equilibrium throughput and social welfare to the operational parameters of the system that can be tuned by a central decision maker, we perform two sets of numerical experiments (comparative statics): (i) in the fraction of time that the system remains unobservable and (ii) in the duration of unobservable periods.

By controlling the fraction of time that the system remains unobservable, we show that typically, a properly adjusted alternating system strictly improves over the continuously observable and continuously unobservable systems in terms of both equilibrium throughput and social welfare. This bridges the two cases that have been considered by Chen and Frank [3], who show that the observable and unobservable systems are preferable for high and low arrival rates respectively. Regulating the fraction γ of informed customers (by tuning the fraction time that the system is observable) has multiple effects on the system. In case of low arrival rates, an increase in γ increases the customer entrance probability for the unobservable periods, but also increases the abandonment probability since the system passes quickly from the unobservable to the observable mode. This double effect is reversed in the case of high arrival rates. If reneging is forbidden and the alternation between observable and unobservable periods becomes very fast, then the present model reduces to the model of Hu, Li and Wang [15]. In this case, our results indicate a unimodal equilibrium throughput which agrees with the findings of [15].

Turning to the second set of experiments, the control of the duration of the unobservable periods can be equivalently viewed as control of the rate of announcements that reveal the queue length to all present customers. The most interesting finding in this case is that the equilibrium throughput and typically also the social welfare are optimized for durations (announcement rates) strictly between 0 and ∞. Decreasing the mean duration of unobservable periods between observable periods of fixed mean length (and hence, increasing the announcement rate) increases both the equilibrium joining probability and the reneging probability. The trade-off between the two effects is not clear and this is the reason for the unimodality of the throughput. Finally, in this setting and in the limiting case of very short observable periods, the present model reduces to the model of Burnetas, Economou and Vasiliadis [2]. The above results are in agreement with the findings of [2] which indicate that the equilibrium throughput is a non-monotonic function of the announcement rate.

A main finding of the numerical experiments concerns the diverse changes in the performance of the system that result from the complex interaction of its operational parameters. Their

conflicting effects on main quantities, such as the joining probability, reneging probability or the expected customer net benefit, preclude general conclusions on equilibrium behavior over broad ranges of parameter values. For practical purposes, this implies that each different sets of parameters should be analyzed individually. However, it also underlines the importance of the derived equilibrium characterization that retains the functionality of the present model and which enables the comparative statics analysis via numerical methods and tools.

We conclude our analysis with a third set of comparative statics on the economic parameters. Specifically, we study the effect on customer strategic behavior of the fraction of the entrance fee that is refundable and of the customers' service valuation. Our results show that the equilibrium throughput is an increasing or unimodal function in the refundable percentage. On the other hand, all performance measures are increasing both in the service valuation and in the service–entrance fee ratio for constant total (entrance and service) fee, as expected.

Our results imply that the alternating information structure of the present model unifies other existing approaches by obtaining as limiting cases both the heterogeneously observable model of [15] and the delayed observations model of [2]. Moreover, it significantly extends them: by fine-tuning the mean durations of the observable and unobservable periods, the present model can typically achieve higher equilibrium throughput and/or social welfare than the corresponding optimal instances of the models in [2] and [15] under the same operational and economic parameters. This comes at no cost in terms of tractability, since the current model always has a unique equilibrium customer strategy that can be characterised via well behaved economic quantities (customer net benefit). Hence, depending on the parameters that can be controlled in a given system or application, the present model is of practical relevance for its optimal design both from a social and a managerial perspective.

1.2 Literature review

Hassin [10] compared the observable and the unobservable versions of the M/M/1 queue with strategic customers regarding their joining/balking dilemma, by focusing on the social welfare and a monopolist's profit under a profit-maximizing admission fee. Let λ and μ denote respectively the arrival and service rate of a M/M/1 queue, R be the service value and C be the waiting cost per time unit for the customers. Hassin showed that if $R\mu \leq 2C$, then the profit under a profit-maximizing admission fee is larger for the observable model, for all $\lambda > 0$. Hence, a profit maximizer prefers to reveal the queue length to the customers. If, however, $R\mu > 2C$, then the profit under a profit-maximizing admission fee is larger for the observable model, if and only if $\lambda \geq \lambda^Z$ for some unique threshold arrival rate, λ^Z. Thus, in this case, a profit maximizer prefers to reveal the queue length only when $\lambda \geq \lambda^Z$. In other words, there is a range of the parameters ($R\mu > 2C$ and $\lambda < \lambda^Z$), for which the provision of more information to the customers hurts the service provider. The same properties also hold for the social welfare under a profit-maximizing fee, but with a different critical value λ^S in place of λ^Z. Thus, there is a range of the parameters ($\lambda < \lambda^S$), for which the provision of more information hurts the society as a whole.

Chen and Frank [3] compared the observable and the unobservable versions of the M/M/1 queue by focusing on the equilibrium effective arrival rate (which is the same as the throughput since there are no abandonments), under an arbitrary fixed admission fee. For a given potential arrival rate λ, let $\lambda_e^{(o)}(\lambda)$ and $\lambda_e^{(u)}(\lambda)$ denote the corresponding equilibrium effective arrival rates in the observable and unobservable versions respectively. Chen and Frank proved that $\lambda_e^{(o)}(\lambda) - \lambda_e^{(u)}(\lambda)$ monotonically increases in λ and there exists a critical value λ^* such that $\lambda_e^{(o)}(\lambda^*) - \lambda_e^{(u)}(\lambda^*) = 0$. Therefore, to attract more customers to the system, it is advisable to conceal the queue length for potential arrival rates λ with $\lambda < \lambda^*$, and to reveal it when $\lambda > \lambda^*$. Shone, Knight and Williams [22] considered the same problem of comparing the equilibrium throughputs, $\lambda_e^{(o)}$ and $\lambda_e^{(u)}$, between the observable and unobservable versions of the M/M/1

queue. They provided necessary and sufficient conditions on the system parameters under which the equilibrium effective arrival rates are equal in the two versions. Moreover, they investigated the behavior of the equilibrium effective arrival rates as functions of the normalized service value $\frac{R\mu}{C}$. In particular, they showed that the number of distinct normalized service values for which $\lambda_e^{(o)} = \lambda_e^{(u)}$ is monotonically increasing with respect to the utilization rate $\rho = \frac{\lambda}{\mu}$ and tends to infinity as $\rho \to 1$.

Guo and Zipkin [8] compared the observable, unobservable and workload observable versions of the M/M/1 queue, under a general reward-cost structure that generalizes the standard Naor's linear reward-cost structure. Under this framework, the service value is R, but a customer's waiting cost is $\theta E[c(W)]$, where W stands for the waiting time, $c(w)$ is a common basic cost function for all customers, and θ is a customer-specific parameter that represents the sensitivity to delay. In other words, a customer with delay sensitivity θ has expected utility $R - \theta E[c(W)]$, if she decides to join. The authors showed that the maximum equilibrium throughput of the system is achieved at different information levels according to the values of the underlying parameters. Wang, Cui and Wang [25] consider an M/M/1 queueing system with a pay-for-priority option, and study customers joint decisions between joining/balking and pay-for-priority. They compare the servers revenue between the observable and the unobservable settings and interestingly, find that the service provider is better off with the observable setting when the system load is either low or high, but benefits more from the unobservable setting when the system load is medium.

The main conclusion from these studies is that the primary factor that determines whether information is good or bad for the service provider and the customers is the distribution function of the customer delay sensitivity and not the common basic cost function. As mentioned above, the relevant literature has reported three main approaches that aim to bridge the observable and unobservable versions of the M/M/1 queue: partially observable models, heterogeneously observable models and models with delayed observations.

In partially observable models, the state-space of the queue length of the M/M/1 queue is partitioned into subsets and the arriving customers are not informed about the exact queue length, but rather about the subset it belongs to. In other words, the waiting space can be considered to be 'compartmented' and the customers are informed only about the compartment in which they are going to be placed. Economou and Kanta [5] considered the case of regular compartmentalization (all compartments being of the same size), and showed that an ideal compartment-size exists, and it may be strictly between 1 (which corresponds to the observable version) and ∞ (which corresponds to the unobservable version). Guo and Zipkin [9] considered the general case of compartments with possibly different sizes and proved several interesting results about the comparison of two partitions of the state-space, one a refinement of the other. More recently, Simhon, Hayel, Starobinski and Zhu [23] considered the M/M/1 queue with strategic customers that face the dilemma of joining/balking when the administrator informs the customers about the current queue length only when it is short, i.e., when it does not exceed a certain threshold D. This corresponds to the partition of the state-space to the subsets $\{0\}, \{1\}, \{2\}, \ldots, \{D\}$ and $\{D+1, D+2, \ldots\}$. The authors proved that the equilibrium throughput is a monotone function of D and hence, if the administrator's goal is to maximize throughput, then the optimal policy is one of the extremes, either the observable or the unobservable queue. Kim and Kim [17] considered the generalization of the last model, by assuming that the customers are informed about the current queue length only when it belongs to an arbitrary subset O. The authors proved the counter-intuitive result that the optimal partition for the maximization of the throughput of the system corresponds to a set O that contains all the states above a threshold, i.e., it is preferable to allow the customers to observe the queue length only when it is large! Finally, Hassin and Koshman [13] considered a model where the arriving customers are only informed about whether the queue length is less than an exogenously given threshold N or not. They focused on the profit maximization problem for the dynamic pricing

version of this model (i.e., different prices are offered to the customers according to whether the queue length is below N or not) and proved the interesting result that the choice of $N = 1$ (meaning that customers are informed only if the server is idle) guarantees at least half of the maximum value that can be generated by the system.

In heterogeneously observable models, the population of customers is divided into informed and uninformed customers, i.e., only a fraction of customers is allowed to observe the queue length before making their decisions. Two such models have been studied by Economou and Grigoriou [4] and Hu, Li and Wang [15]. Economou and Grigoriou determined the equilibrium strategies in the case where the service values and the waiting costs are different for informed and uninformed customers. Hu, Li and Wang proved that throughput and social welfare are in general unimodal and not monotonous in the fraction of informed customers. In other words, information heterogeneity in the population can lead to more efficient outcomes in terms of the system throughput or social welfare than information homogeneity. Moreover, they showed that for an overloaded system (with utilization factor sufficiently higher than 1), social welfare always reaches its maximum when some fraction of customers is uninformed.

In models with delayed observations, the customers decide whether to join or balk without knowing the state of the system, but later on they are informed about their current position and may renege. Burnetas, Economou and Vasiliadis [2] considered the M/M/1 queue where the administrator of the system makes periodic announcements to the customers about their current positions. The model was motivated by a situation that occurs when people submit petitions through certain web-based systems. Then, upon submission, the customers receive a confirmation message with the registration number of their petition. Later on, they learn the number of pending petitions in front of them. This is done either by periodic refreshments of a web page that indicates the registration number of the currently processed petition or by periodic bulk emails that announce the status of the pending petitions. The authors have shown that the equilibrium throughput is a non-monotonic function of the announcement rate. This implies that there exists an ideal announcement rate, strictly between 0 and ∞, that maximizes the throughput. In other words, some delay in providing information to the customers is beneficial in terms of throughput. Another model with delayed observation characteristics is the so-called 'armchair decision' problem introduced by Hassin and Roet-Green [14] (see also Roet-Green [21]). In this model, the customers observe the queue length before reaching it, using probably some web-based application. Then, they decide whether to leave their armchairs and go to the service facility or not, but when they arrive at the system, they are informed about the current queue length and should make their second decision, to join or balk. For more papers and thorough overviews of the results that concern the control of information in queueing systems with strategic customers, see Section 3.5, *Information Control*, in Hassin [11] and the recent review paper of Ibrahim [16].

The rest of the paper is organized as follows: In Section 2, we define the model with the underlying reward-cost structure and derive sufficient conditions for customer equilibrium strategies. Section 3 presents our computations on system performance that lead to the characterization of equilibrium customer strategies in Section 4. In Section 5, we perform a number of numerical experiments and conclude our analysis with some useful take-away messages and managerial insight in Section 6. Some technical material is presented in Appendix A.

2   Model and strategies

We consider an M/M/1 queue with arrival rate λ and service rate μ, where arriving customers decide whether to join or balk. The system alternates between unobservable and observable periods that are exponentially distributed with rates θ and ζ, respectively. The customers are homogeneous in their valuations. More specifically, each one of them values service R units and accumulates costs at rate C, as long as she stays in the system, either waiting or being served.

6

The administrator of the system imposes an entrance fee f_e, which is paid by each customer who decides to join. Moreover, he imposes a service fee f_s which is paid only by those customers that receive service. There is also a refund r which is given to customers that decide to renege before their service has been completed. The refund may be positive (partial or full refund of the entrance fee) or negative (which means that there is a penalty for an abandonment) or even $-\infty$ (meaning that reneging is impossible or forbidden).

We assume that the customers are strategic and risk neutral, in the sense that they want to maximize their expected net benefit, knowing that the others have similar objectives. During observable periods, the customers are informed about the queue length upon arrival and then make their joining/balking decisions, whereas in unobservable periods, they make their joining/balking decisions relying solely on their knowledge of the system parameters. All customers are informed about their current positions when the system enters an observable period and any time, they may decide to renege.

To avoid trivial cases, we assume throughout the paper that the following two conditions hold:

$$R > f_e + f_s + \frac{C}{\mu}, \tag{2.1}$$

and

$$r \le f_e. \tag{2.2}$$

Condition (2.1) ensures that the value of service is high enough so that a customer that observes an empty system prefers to join; otherwise the system would be continuously empty. Condition (2.2) ensures that there are no customers that enter and remain in the system only instantaneously. Indeed, if $r > f_e$, then even a customer who observes a huge queue length upon arrival is willing to enter, but will renege immediately.

To analyze customer strategic behavior, we should take into account that the arriving customers face the dilemma of whether to join or balk upon arrival, and then, the joining customers continuously face the dilemma of whether to stay or renege.

If a customer arrives at an observable period and finds n customers in the system, then she enters if and only if $R - f_e - f_s - C\frac{n+1}{\mu} \ge 0$ or equivalently if

$$n + 1 \le \left\lfloor \frac{\mu(R - f_e - f_s)}{C} \right\rfloor = n_e. \tag{2.3}$$

Hence such a customer uses Naor's threshold and enters if her position in the system after joining is at most n_e, given by (2.3). If she enters, it is certain that she will stay in the system till her service completion, because no matter what the other customers do, her expected net benefit will be non-negative at all subsequent moments (because of the FCFS discipline and the Markovian framework).

If a customer arrives at an unobservable period, then she will enter with some probability q. If she enters, then she will certainly stay till the first time that the system becomes observable or till her service completion, whatever occurs first (again because of the assumption of exponentially distributed times). Suppose that the system becomes observable before the service completion of a customer that joined in an unobservable period. If n is the current position of the customer, then she will stay if and only if $R - f_s - C\frac{n}{\mu} \ge r$ or equivalently if

$$n \le \left\lfloor \frac{\mu(R - r - f_s)}{C} \right\rfloor = n_s. \tag{2.4}$$

Note that f_e does not play any role for this decision, since it is refundable only to the extent specified by r. Of course, if the tagged customer decides to stay after the moment that the system becomes observable, then she will stay till her service completion, as her expected benefit cannot deteriorate. Note, also, that

$$n_e \le n_s$$

7

Economic parameters	
$R > 0$	customer service valuation
$C > 0$	customer waiting cost per time unit in the system
$f_e \geq 0$	entrance fee
$f_s \geq 0$	service fee
r	$r \geq 0$: refund to reneging customers
	$r < 0$: penalty for abandonment
	$r = -\infty$: reneging is forbidden

Operational parameters	
mode 1	observable system
mode 0	unobservable system
$\lambda > 0$	arrival rate
$\mu > 0$	service rate
$\theta > 0$	exponential rate of the duration of unobservable periods
$\zeta > 0$	exponential rate of the duration of observable periods
$B > 0$	mean duration of an information cycle: $B = 1/\zeta + 1/\theta$
$\gamma \in (0,1)$	fraction of customers who arrive at observable mode: $\gamma = \theta/(\zeta + \theta)$
(n_e, n_s, q)	potential equilibrium strategy (PES)
$n_e \geq 0$	join/balk threshold (Naor's) at arrival instants during observable mode
$n_s \geq 0$	stay/renege threshold at change instants from unobservable to observable mode
$q \in [0,1]$	joining probability in unobservable mode

Performance measures	
$q_e \in [0,1]$	equilibrium joining probability in unobservable mode
$\mu_e \geq 0$	equilibrium throughput: mean number of service completions per time unit
$S_e \geq 0$	equilibrium social welfare per time unit

Table 1: Model parameters and notation

because of condition (2.2). From the above discussion, we conclude that an equilibrium strategy should

- prescribe *enter* if the offered position for an arriving customer during an observable period is at most n_e given by (2.3),
- prescribe *enter* with some probability q if a customer arrives during an unobservable period, and
- prescribe *stay* if the position of a customer at the time of a change from an unobservable to observable mode is at most n_s given by (2.4).

Thus, a *potential equilibrium strategy* (PES) should satisfy the above three conditions and will be referred to as an (n_e, n_s, q)-PES. Note that only q remains to be determined. All model parameters are summarized for convenience in Table 1.

3   Customer expected net benefit

Suppose that the population of the customers adopts an (n_e, n_s, q)-PES and consider a tagged customer. Her best response against the (n_e, n_s, q)-PES will necessarily be an (n_e, n_s, q')-PES, with a possibly different joining probability q'. This follows from the discussion in Section 2. To compute the best response of the tagged customer, suppose that she arrives during an

unobservable period and let $\mathcal{U}(n_e, n_s, q)$ be her expected net benefit if the other customers follow the (n_e, n_s, q)-PES and the tagged customer decides to join and will use the same threshold n_s for staying/reneging as the other customers at the time where the system becomes observable.

To determine $\mathcal{U}(n_e, n_s, q)$, we condition on the number of customers N_q^- that are present in the system just before the arrival of the tagged customer (noting that N_q^- is not observable for her). Then,

$$\mathcal{U}(n_e, n_s, q) = \sum_{n=0}^{\infty} \Pr[N_q^- = n] U(n; n_s), \tag{3.1}$$

where $U(n; n_s)$ stands for the conditional expected net benefit of the tagged customer if she decides to join, given that n customers are present in the system upon her arrival (excluding herself) and all of them use the same reneging threshold n_s. Note that $U(n; n_s)$ is independent from the strategy parameters n_e, q, and the operational model parameters λ, ζ. Also, the subscript q in N_q^- is present to emphasize the dependence on q which is crucial in several derivations. To evaluate $\mathcal{U}(n_e, n_s, q)$ using (3.1), we start with computing the conditional expected values $U(n; n_s)$.

Theorem 3.1. *The conditional expected net benefit of an arriving customer if she decides to join, given that n customers are present in the system upon her arrival and all of them use the reneging threshold n_s is given by*

$$U(n; n_s) = \begin{cases} R - f_e - f_s - \frac{C(n+1)}{\mu} & \text{if } 0 \le n < n_s \\ r - f_e - \frac{C}{\theta} + (R - r - f_s - \frac{Cn_s}{\mu} + \frac{C}{\theta})(\frac{\mu}{\mu+\theta})^{n+1-n_s} & \text{if } n \ge n_s. \end{cases} \tag{3.2}$$

Proof. Consider a tagged customer that arrives and decides to join during an unobservable period. Suppose that the system has n other customers at that moment. We consider two cases according to whether $n < n_s$ or not.

In the first case, where $n < n_s$, the customer will certainly complete her service. Therefore, she will receive the service reward R and will pay the entrance and service fees, f_e and f_s. Moreover, her sojourn time in the system will be the sum of $n+1$ service times and we conclude with the first branch of (3.2).

In the second case, where the tagged customer arrives when there are $n \ge n_s$ customers and decides to join, let X be the time till the beginning of the first observable period after her arrival and K be the number of service completions that occur during X. Let U_n be the net benefit of the tagged customer. Then, by conditioning on X and K we have

$$U(n; n_s) = \int_0^{\infty} \sum_{k=0}^{\infty} E[U_n \mid X = x, K = k] \Pr[K = k \mid X = x] f_X(x) dx, \tag{3.3}$$

where $f_X(x)$ is the probability density of X. Note that $f_X(x) = \theta e^{-\theta x}$, $x > 0$, because of the memoryless property of the exponential distribution, and that $\Pr[K = k \mid X = x] = e^{-\mu x} \frac{(\mu x)^k}{k!}$, $k \ge 0$, assuming that the server is always busy during time X. Moreover, the conditional mean value $E[U_n \mid X = x, K = k]$ is given by

$$E[U_n \mid X = x, K = k] = \begin{cases} -f_e - Cx + r, & \text{if } 0 \le k \le n - n_s, \\ -f_e - C\left(x + \frac{n+1-k}{\mu}\right) + R - f_s, & \text{if } n - n_s + 1 \le k \le n, \\ -f_e - C\frac{(n+1)x}{k+1} + R - f_s, & \text{if } k \ge n + 1. \end{cases} \tag{3.4}$$

Indeed, to justify (3.4), we notice that there are three cases regarding the tagged customer:

Case I: At the beginning of the first observable period after her arrival, the tagged customer occupies a position greater than n_s.

9

In this case, the number of service completions k is such that $n + 1 - k \geq n_s + 1$, i.e., $k \leq n - n_s$. The tagged customer will renege at time x, so her sojourn time in the system is the time till the first announcement x. Therefore, she will pay f_e upon entrance, suffer waiting cost Cx and receive the refund r upon abandonment. So, we conclude the first branch of (3.4).

Case II: At the beginning of the first observable period after her arrival, the tagged customer occupies a position less than or equal to n_s.

In this case, the number of service completions k is such that $1 \leq n + 1 - k \leq n_s$, i.e., $n - n_s + 1 \leq k \leq n$. The tagged customer will not renege and has to wait for the completion of $n + 1 - k$ service times after the beginning of the observable period that followed her arrival. Her mean sojourn time in the system will be $x + \frac{n+1-k}{\mu}$. Therefore, she will suffer waiting cost $C(x + \frac{n+1-k}{\mu})$ and the net service value will be $R - f_e - f_s$, so we conclude the second branch of (3.4).

Case III: At the beginning of the first observable period after her arrival, the tagged customer has already been served.

In this case, the number of service completions k is such that $k \geq n + 1$. In the interval of length x, till the beginning of the observable period, there were k events in the Poisson process of the service completions and the departure of the tagged customer corresponds to the $(n + 1)$-th event of these k events. But then, the conditional distribution of the departute time of the tagged customer given x coincides with the $(n + 1)$-order statistic of k i.i.d. uniform random variables in $[0, x]$ (see e.g., Campbell's Theorem in Section 5.3 of Kulkarni [18]). Its mean value is $\frac{(n+1)x}{k+1}$, so the waiting cost for the tagged customer is $C \frac{(n+1)x}{k+1}$, while the net service value is $R - f_e - f_s$. We conclude with the third branch of (3.4).

We can now insert (3.4) into (3.3) and we obtain

$$U(n; n_s) = \int_0^\infty \sum_{k=0}^{n-n_s} (-f_e - Cx + r) e^{-\mu x} \frac{(\mu x)^k}{k!} \theta e^{-\theta x} dx$$

$$+ \int_0^\infty \sum_{k=n-n_s+1}^{n} (-f_e - Cx - C \frac{n+1-k}{\mu} + R - f_s) e^{-\mu x} \frac{(\mu x)^k}{k!} \theta e^{-\theta x} dx$$

$$+ \int_0^\infty \sum_{k=n+1}^{\infty} (-f_e - C \frac{(n+1)x}{k+1} + R - f_s) e^{-\mu x} \frac{(\mu x)^k}{k!} \theta e^{-\theta x} dx.$$

Evaluating the integrals (using the formula $\int_0^\infty x^h e^{-\nu x} dx = \frac{h!}{\nu^{h+1}}$) and grouping similar terms yields

$$U(n; n_s) = - f_e + r \sum_{k=0}^{n-n_s} \frac{\mu^k \theta}{(\mu + \theta)^{k+1}} + (R - f_s) \sum_{k=n-n_s+1}^{\infty} \frac{\mu^k \theta}{(\mu + \theta)^{k+1}}$$

$$- C \frac{1}{\mu} \sum_{k=1}^{n-n_s} \frac{k \mu^k \theta}{(\mu + \theta)^{k+1}} - C \frac{n+1}{\mu} \sum_{k=n-n_s+1}^{\infty} \frac{\mu^k \theta}{(\mu + \theta)^{k+1}}, \ n \geq n_s.$$

Evaluating the geometric sums and grouping equal terms yields after some simplifications the second branch of (3.2). □

To proceed with the computation of $\mathcal{U}(n_e, n_s, q)$ using (3.1), we need to compute the probabilities $\Pr[N_q^- = n]$. These probabilities can be computed by studying the dynamics of the system, when the population of customers follow the (n_e, n_s, q)-PES. Indeed, under this strategy, the state of the system is described by a continuous-time Markov chain (CTMC) $\{(N(t), I(t)) :$

$t \geq 0\}$, where $N(t)$ records the number of customers in the system at time t, while $I(t)$ denotes the mode of operation ($I(t) = 1$ during observable periods and $I(t) = 0$ during unobservable periods). The state-space of $\{(N(t), I(t))\}$ is $\mathcal{S}_{N,I} = \{(n,0) : n \geq 0\} \cup \{(n,1) : 0 \leq n \leq n_s\}$ and the corresponding transition diagram is shown in Figure 1.
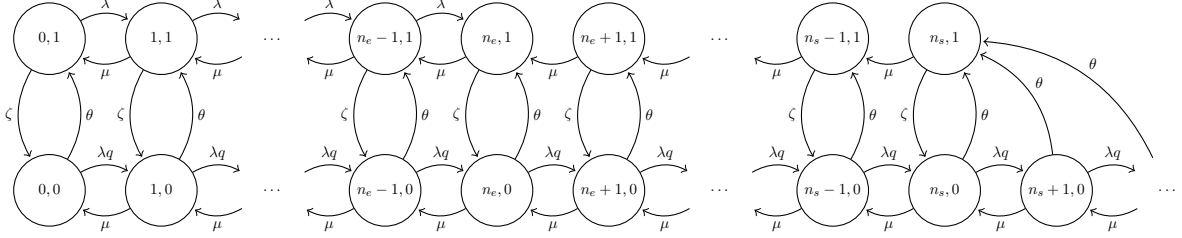

Figure 1: Transition diagram of the system state, when the customers follow the (n_e, n_s, q)-PES.

Its transition rates are given by

$$
q_{(n,i)(m,j)} = \begin{cases}
\lambda q & \text{if } i = j = 0, \quad n \geq 0, \;\; m = n+1, \\
\mu & \text{if } i = j = 0, \quad n \geq 1, \;\; m = n-1, \\
\lambda & \text{if } i = j = 1, \quad 0 \leq n \leq n_e - 1, \;\; m = n+1, \\
\mu & \text{if } i = j = 1, \quad 1 \leq n \leq n_s, \;\; m = n-1, \\
\theta & \text{if } i = 0, j = 1, \quad 0 \leq n \leq n_s - 1, \;\; m = n, \\
\theta & \text{if } i = 0, j = 1, \quad n \geq n_s, \;\; m = n_s, \\
\zeta & \text{if } i = 1, j = 0, \quad 0 \leq n \leq n_s, \;\; m = n, \\
0 & \text{otherwise.}
\end{cases}
$$

Denote by $(p(n,i) : (n,i) \in \mathcal{S}_{N,I})$ its steady-state distribution. Then,

$$
\Pr[N_q^- = n] = \frac{p(n,0)}{\sum_{m=0}^{\infty} p(m,0)}. \tag{3.5}
$$

Therefore, we need to compute $(p(n,i) : (n,i) \in \mathcal{S}_{N,I})$. The distribution $(p(n,i))$ is the unique normalized solution of the following balance equations:

$$
(\lambda + \zeta)p(0,1) = \theta p(0,0) + \mu p(1,1), \tag{3.6}
$$
$$
(\lambda + \mu + \zeta)p(n,1) = \lambda p(n-1,1) + \theta p(n,0) + \mu p(n+1,1), \quad 1 \leq n \leq n_e - 1, \tag{3.7}
$$
$$
(\mu + \zeta)p(n_e,1) = \lambda p(n_e - 1, 1) + \theta p(n_e, 0) + \mu p(n_e + 1, 1), \tag{3.8}
$$
$$
(\mu + \zeta)p(n,1) = \theta p(n,0) + \mu p(n+1,1), \quad n_e + 1 \leq n \leq n_s - 1, \tag{3.9}
$$
$$
(\mu + \zeta)p(n_s,1) = \theta \sum_{n=n_s}^{\infty} p(n,0), \tag{3.10}
$$
$$
(\lambda q + \theta)p(0,0) = \zeta p(0,1) + \mu p(1,0), \tag{3.11}
$$
$$
(\lambda q + \mu + \theta)p(n,0) = \lambda q p(n-1,0) + \zeta p(n,1) + \mu p(n+1,0), \quad 1 \leq n \leq n_s, \tag{3.12}
$$
$$
(\lambda q + \mu + \theta)p(n,0) = \lambda q p(n-1,0) + \mu p(n+1,0), \quad n \geq n_s + 1. \tag{3.13}
$$

Solving the system of the balance equations and the normalization equation

$$
\sum_{(n,i)\in\mathcal{S}_{N,I}} p(n,i) = 1, \tag{3.14}
$$

and using (3.1), (3.2) and (3.5), we can compute $\mathcal{U}(n_e, n_s, q)$ which is the key quantity for deriving the customer equilibrium behavior. The system of the balance equations and the normalization equation is quite involved. However, it can be reduced easily to a finite linear

11

system because of the homogeneous 1-dimensional nature of the model for $n \geq n_s + 1$. Indeed, (3.13) is a homogeneous linear second-order difference equation. Therefore, its solution is

$$p(n, 0) = c_+ \rho_+^n + c_- \rho_-^n, \quad n \geq n_s, \tag{3.15}$$

where ρ_+ and ρ_- are the roots of the corresponding characteristic equation

$$\mu x^2 - (\lambda q + \mu + \theta)x + \lambda q = 0 \tag{3.16}$$

and c_+, c_- constants to be determined (see e.g., Section 2.3 in Elaydi [7]). Therefore,

$$\rho_{+,-} = \frac{\lambda q + \mu + \theta \pm \sqrt{(\lambda q + \mu + \theta)^2 - 4\lambda q \mu}}{2\mu}. \tag{3.17}$$

It is easy now to see that $\rho_+ > 1$, whereas $0 < \rho_- < 1$. Because of the normalization equation, $\sum_{n=n_s}^{\infty} p(n, 0) < \infty$, so necessarily the coefficient c_+ of ρ_+ in (3.15) should be 0. Hence, (3.15) becomes

$$p(n, 0) = c_- \rho_-^n = \left(\frac{\lambda q + \mu + \theta - \sqrt{(\lambda q + \mu + \theta)^2 - 4\lambda q \mu}}{2\mu} \right)^n p(n_s, 0), \quad n \geq n_s. \tag{3.18}$$

Now, equation (3.10) is written as

$$(\mu + \zeta)p(n_s, 1) = \frac{\theta}{1 - \rho_-} p(n_s, 0), \tag{3.19}$$

and equation (3.12) for $n = n_s$ reduces to

$$(\lambda q + \mu + \theta)p(n_s, 0) = \lambda q p(n_s - 1, 0) + \zeta p(n_s, 1) + \mu \rho_- p(n_s, 0).$$

The latter can be also written as

$$(\frac{\theta}{1 - \rho_-} + \mu)p(n_s, 0) = \zeta p(n_s, 1) + \lambda q p(n_s - 1, 0), \tag{3.20}$$

using that ρ_- satisfies (3.16). Now, equations (3.6)-(3.9), (3.19), (3.11), (3.12) for $1 \leq n \leq n_s - 1$ and (3.20) show that $(p(n, i) : i = 0, 1$ and $0 \leq n \leq n_s)$ satisfies the balance equations of the finite non-homogeneous Quasi-Birth-Death (QBD) process that results from the original chain in Figure 1, when the states $(n, 0)$ for $n \geq n_s + 1$ are omitted and the rate from $(n_s, 0)$ to $(n_s, 1)$ becomes $\theta/(1 - \rho_-)$. Indeed, this finite QBD process, with transition rate diagram given in Figure 2, is the censored process that results from the original chain observed only while it stays in states with $n \leq n_s$ (for details see [19], Chapter 5, in particular Section 5.5). Hence, $(p(n, i) : i = 0, 1$ and $0 \leq n \leq n_s)$ can be effectively computed up to a normalization constant by using any general algorithm for the computation of the steady-state distributions of finite QBD processes (see e.g., [19], Chapter 12 or [1], Section 7.2.1). Then, the remaining steady-state probabilities $p(n, 0)$, $n \geq n_s + 1$, are computed up to the same normalization constant by (3.18). Finally, the normalization constant is computed using the normalization equation (3.14). Note that the algorithms for the computation of the steady-state probabilities of a finite QBD are very fast since they are of block-Gaussian elimination type that exploit the block-tridiagonal form of the transition rate matrix.

We now proceed to obtain a formula for $\mathcal{U}(n_e, n_s, q)$. By (3.1) and (3.5), we have that

$$\mathcal{U}(n_e, n_s, q) = \sum_{n=0}^{\infty} \frac{p(n, 0)}{\sum_{m=0}^{\infty} p(m, 0)} U(n; n_s) = \frac{\zeta + \theta}{\zeta} \sum_{n=0}^{\infty} p(n, 0)U(n; n_s), \tag{3.21}$$
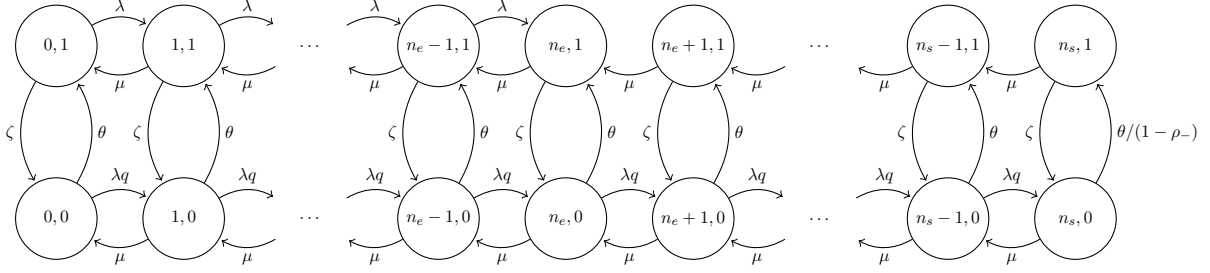
Figure 2: Transition diagram of the censored process in the set of states with $n \le n_s$.

because $\{I(t)\}$ is a 2-state CTMC with rate $q_{01} = \theta$ and $q_{10} = \zeta$ and so we conclude that $\sum_{m=0}^{\infty} p(m,0) = \frac{\zeta}{\zeta+\theta}$. To evaluate the sum in (3.21), we decompose it in two sums according to the two branches of (3.2). We have:

$$\sum_{n=0}^{n_s-1} p(n,0)U(n;n_s) = \sum_{n=0}^{n_s-1} p(n,0)\left(R - f_e - f_s - \frac{C(n+1)}{\mu}\right), \tag{3.22}$$

$$\sum_{n=n_s}^{\infty} p(n,0)U(n;n_s) = \sum_{n=n_s}^{\infty} p(n,0)\left(r_a - f_e - \frac{C}{\theta}\right)$$

$$+ \sum_{n=n_s}^{\infty} p(n,0)\left(R - r_a - f_s - \frac{Cn_s}{\mu} + \frac{C}{\theta}\right)\left(\frac{\mu}{\mu+\theta}\right)^{n+1-n_s}. \tag{3.23}$$

The right-hand sides of (3.22)-(3.23) can be written more compactly in terms of the following partial generating functions of the steady-state distribution $(p(n,i) : (n,i) \in \mathcal{S}_{N,I})$, that correspond to the various groups of states that appear in the transition diagram (except from the state $(n_s,1)$ which forms a group by itself):

$$P_{0a}(z) = \sum_{n=0}^{n_e-1} p(n,0)z^n, \quad P_{0b}(z) = \sum_{n=n_e}^{n_s-1} p(n,0)z^{n-n_e}, \quad P_{0c}(z) = \sum_{n=n_s}^{\infty} p(n,0)z^{n-n_s} \tag{3.24}$$

$$P_{1a}(z) = \sum_{n=0}^{n_e-1} p(n,1)z^n, \quad P_{1b}(z) = \sum_{n=n_e}^{n_s-1} p(n,1)z^{n-n_e}. \tag{3.25}$$

Indeed, equations (3.22)-(3.23) assume the form

$$\sum_{n=0}^{n_s-1} p(n,0)U(n;n_s) = \left(R - f_e - f_s - \frac{C}{\mu}\right)P_{0a}(1) - \frac{C}{\mu}P'_{0a}(1)$$

$$+ \left(R - f_e - f_s - \frac{C}{\mu}\right)P_{0b}(1) - \frac{C}{\mu}P'_{0b}(1) - \frac{Cn_e}{\mu}P_{0b}(1), \tag{3.26}$$

$$\sum_{n=n_s}^{\infty} p(n,0)U(n;n_s) = \left(r_a - f_e - \frac{C}{\theta}\right)P_{0c}(1)$$

$$+ \left(R - r_a - f_s - \frac{Cn_s}{\mu} + \frac{C}{\theta}\right)\frac{\mu}{\mu+\theta}P_{0c}\left(\frac{\mu}{\mu+\theta}\right). \tag{3.27}$$

Combining (3.21) with (3.26)-(3.27) yields an expression for $\mathcal{U}(n_e, n_s, q)$ which is reported in the following Theorem.

Theorem 3.2. *The unconditional expected net benefit of an arriving customer if she decides to join, given that the population of the customers follow the (n_e, n_s, q)-PES is given by*

$$\mathcal{U}(n_e, n_s, q) = \frac{\zeta + \theta}{\zeta} \left[\left(R - f_e - f_s - \frac{C}{\mu} \right) \left(P_{0a}(1) + P_{0b}(1) \right) - \frac{Cn_e}{\mu} P_{0b}(1) \right.$$
$$+ \left(r_a - f_e - \frac{C}{\theta} \right) P_{0c}(1) - \frac{C}{\mu} \left(P'_{0a}(1) + P'_{0b}(1) \right)$$
$$\left. + \left(R - r_a - f_s - \frac{Cn_s}{\mu} + \frac{C}{\theta} \right) \frac{\mu}{\mu + \theta} P_{0c} \left(\frac{\mu}{\mu + \theta} \right) \right].$$

The probability generating functions that are needed to evaluate $\mathcal{U}(n_e, n_s, q)$ can be computed either using the steady-state probabilities that are obtained via the finite QBD approach that we described above, or directly using a generating function approach. The latter approach is interesting and efficient, but quite involved. So we describe it in detail in the Appendix of the paper. Because of its complexity, it may be preferable than the former approach only when n_s is large, in which case the finite QBD approach is computationally costly. Having determined the unconditional expected net benefit function $\mathcal{U}(n_e, n_s, q)$, we can now proceed towards the characterization and computation of the equilibrium strategies.

4 Monotonicity properties and equilibrium behavior

In this section, we characterize the equilibrium customer behavior using the performance evaluation results for the model under an (n_e, n_s, q)-PES that were reported in Section 3. We first derive several monotonicity properties of the model associated with the functions $U(n; n_s)$ and $\mathcal{U}(n_e, n_s, q)$ which are crucial for the study of the equilibrium customer behavior.

Proposition 4.1. *The steady-state number, N_q^-, of customers in the system at arrival instants during unobservable periods, when the (n_e, n_s, q)-PES is followed by the customer population, is stochastically increasing in q.*

Proof. Consider two systems, 1 and 2, with identical operational parameters λ, μ, θ and ζ, and identical economic parameters R and C, where the customers have adopted the thresholds n_e and n_s given by (2.3) and (2.4), respectively. The two systems differ only in the join probability q when customers arrive during an unobservable period. More concretely, we suppose that the customers enter with probability $q^{(i)}$ when they arrive at an unobservable period of system i, for $i = 1, 2$. Suppose that $q^{(1)} < q^{(2)}$ and let $\{(N^{(i)}(t), C^{(i)}(t))\}$ be the CTMC describing system i, for $i = 1, 2$. We construct a coupling of the two processes that represent the states of the two systems as follows:

The observable and unobservable periods alternate identically in the two systems. The service completions are identical in the two systems and are generated by the same Poisson process $\{M(t)\}$ with rate μ (when any of the two systems is empty, the Poisson generated events do not have any influence on the corresponding state 0 of $\{N^{(i)}(t)\}$). The arrivals at system 2 are generated by a Poisson process $\{\Lambda^{(2)}(t)\}$ with rate $\lambda q^{(2)}$. On the other hand, for system 1, we assume that an arrival occurs at an event of the Poisson process $\{\Lambda^{(2)}(t)\}$ with probability $q^{(1)}/q^{(2)}$. This ensures that the arrivals at system 1 constitute a Poisson process $\{\Lambda^{(1)}(t)\}$ with rate $\lambda q^{(1)}$.

A comparison of the coupled realizations of the two processes $\{N^{(1)}(t)\}$ and $\{N^{(2)}(t)\}$ shows that both processes move one step to the left, when an event at the service completion Poisson process $\{M(t)\}$ occurs. Moreover, whenever a change happens in the informational process, from the observable to the unobservable mode, whichever of the processes $\{N^{(1)}(t)\}$ and $\{N^{(2)}(t)\}$ are above n_s at the change instant moves to state n_s. If only one of them is above n_s, then only this process is influenced. Finally, when an event of the Poisson process $\{\Lambda^{(2)}(t)\}$ occurs,

the process $\{N^{(2)}(t)\}$ certainly moves one step to the right, while the process $\{N^{(1)}(t)\}$ moves one step to the right with probability $q^{(1)}/q^{(2)}$. Therefore, a moment of reflection shows that if the processes $\{N^{(1)}(t)\}$ and $\{N^{(2)}(t)\}$ start from $n^{(1)}$ and $n^{(2)}$ customers respectively with $n^{(1)} \leq n^{(2)}$, then the sample-path of $\{N^{(1)}(t)\}$ remains 'under' the corresponding sample-path of $\{N^{(2)}(t)\}$ for all t. This proves that $\{N^{(1)}(t)\} \leq_{st} \{N^{(2)}(t)\}$. In particular, if we consider the two systems only during their unobservable periods, we conclude that the sample-path of $\{N^{(1)}(t)\}$ remains 'under' the corresponding sample-path of $\{N^{(2)}(t)\}$ so the corresponding steady-state distributions are also stochastically ordered. But during unobservable periods, the arrival processes at both systems are Poisson, so the steady-state distributions of the number of customers in continuous time and at arrival instants coincide (by applying the Poisson-Arrivals-See-Time-Averages (PASTA) result). Therefore, if we denote by $N_{q^{(i)}}^-$ the steady-state distribution of the number of customers in system i at arrival instants when the system is unobservable, $i = 1, 2$, we conclude that $N_{q^{(1)}}^- \leq_{st} N_{q^{(2)}}^-$. $\qquad\square$

Proposition 4.2. *The conditional expected net benefit function $U(n; n_s)$ is strictly decreasing in n, for any fixed reneging threshold n_s.*

Proof. Proof The top branch of $U(n; n_s)$ of (3.2) is obviously strictly decreasing in n. The bottom branch is also strictly decreasing in n. Indeed, by the definition of n_s (see (2.4)), we have that $R - f_s - C\frac{n_s}{\mu} \geq r$, so the coefficient $R - r - f_s - \frac{Cn_s}{\mu} + \frac{C}{\theta}$ in the bottom branch of (3.2) is positive. Moreover, $(\frac{\mu}{\mu+\theta})^{n+1-n_e}$ is strictly decreasing in n.

It remains to show that $U(n; n_s)$ remains strictly decreasing at its turning point from the top branch to the bottom, i.e., that

$$R - f_e - f_s - \frac{Cn_s}{\mu} > r - f_e - \frac{C}{\theta} + (R - r - f_s - \frac{Cn_s}{\mu} + \frac{C}{\theta})\frac{\mu}{\mu+\theta}.$$

After some simplification, this is equivalently written as $R - r - f_s - \frac{Cn_s}{\mu} + \frac{C}{\theta} > 0$ which is valid by the definition of n_s. $\qquad\square$

Proposition 4.3. *The unconditional expected net benefit function $\mathcal{U}(n_e, n_s, q)$ is strictly decreasing in q, for fixed thresholds n_e and n_s.*

Proof. For fixed n_s, we have that $U(n; n_s)$ does not depend on q, nor on n_e, so we can write (3.1) as

$$\mathcal{U}(n_e, n_s, q) = E[U(N_q^-; n_s)],$$

where N_q^- is the random variable that was defined in the statement of Proposition 4.1. Now, $U(n; n_s)$ is strictly decreasing in n because of Proposition 4.2 and N_q^- is stochastically increasing in q by Proposition 4.1. Therefore, $q_1 < q_2$ implies that $\mathcal{U}(n_e, n_s, q_1) = E[U(N_{q_1}^-; n_s)] > E[U(N_{q_2}^-; n_s)] = \mathcal{U}(n_e, n_s, q_2)$. $\qquad\square$

We are now ready to state and prove the characterization of equilibrium customer strategies.

Theorem 4.1. *An equilibrium strategy always exists and is unique. It is the (n_e, n_s, q_e)-PES with n_e, n_s and q_e given respectively from (2.3), (2.4) and*

$$q_e = \begin{cases} 0 & if \, \mathcal{U}(n_e, n_s, 0) \leq 0, \\ q_e^* & if \, \mathcal{U}(n_e, n_s, 1) < 0 < \mathcal{U}(n_e, n_s, 0), \\ 1 & if \, \mathcal{U}(n_e, n_s, 1) \geq 0, \end{cases}$$

where q_e^* is the root of the equation $\mathcal{U}(n_e, n_s, q) = 0$ with respect to q in $(0, 1)$ (which exists and is unique when $\mathcal{U}(n_e, n_s, 1) < 0 < \mathcal{U}(n_e, n_s, 0)$).

Proof. Proof We have already shown during the discussion in Section 2, that an equilibrium strategy is necessarily an (n_e, n_s, q)-PES with n_e and n_s given by (2.3) and (2.4), respectively. Assume that the population of customers follows the (n_e, n_s, q)-PES and consider a tagged customer. We have the following three cases:

Case I: $q = 0$.

    Then, the best response of the tagged customer against $(n_e, n_s, 0)$ is the same strategy, if and only if, the customer prefers to balk if she finds the system at the unobservable mode upon arrival. Therefore, the $(n_e, n_s, 0)$-PES is an equilibrium strategy if and only if $\mathcal{U}(n_e, n_s, 0) \leq 0$.

Case II: $q \in (0, 1)$.

    Then, the best response of the tagged customer against (n_e, n_s, q) is the same strategy, if and only if, the customer is indifferent between joining and balking if she finds the system at the unobservable mode upon arrival. Therefore, the (n_e, n_s, q)-PES is an equilibrium strategy if and only if $\mathcal{U}(n_e, n_s, q) = 0$. However, because of the strict monotonicity of $\mathcal{U}(n_e, n_s, q)$ with respect to q (by Proposition 4.3), we have that the equation $\mathcal{U}(n_e, n_s, q) = 0$ has a solution in $(0, 1)$ if and only if $\mathcal{U}(n_e, n_s, 1) < 0 < \mathcal{U}(n_e, n_s, 0)$ and in this case, it is unique.

Case III: $q = 1$.

    Then, the best response of the tagged customer against $(n_e, n_s, 1)$ is the same strategy, if and only if, the customer prefers to join if she finds the system at the unobservable mode upon arrival. Therefore, the $(n_e, n_s, 1)$-PES is an equilibrium strategy if and only if $\mathcal{U}(n_e, n_s, 1) \geq 0$.

<div align="right">□</div>

Theorem 4.1 has far-reaching implications in terms of the economic and operational analysis of the alternating information structure. By establishing existence of a unique equilibrium and by characterizing it in terms of the well behaved customer expected net benefit (cf. Theorem 3.2), it essentially states that the current model remains tractable despite its increased complexity over other benchmark approaches, [15, 2]. This structure can be utilized to study the model's performance in equilibrium via numerical comparative statics on its operational and economic parameters.

5   Effects of the alternating observational structure: Numerical experiments

As previously mentioned, the alternation between observable and unobservable periods represents a number of different situations that arise in practice. These situations determine which parameters can be controlled by a decision maker and hence, the ways in which system performance can be improved. Our objective is to study such effects on strategic customer behavior and to derive managerial insight on optimizing system design.

Specifically, we are interested in the behavior of the equilibrium joining probability q_e, the equilibrium throughput of the system $\mu_e = \mu(1 - p(0, 0) - p(0, 1))$ (i.e., the number of service completions per time unit) and the equilibrium social welfare

$$S_e = R\mu_e + ra_e - CE[N_{q_e}],$$

where $a_e = \sum_{n=n_s+1}^{\infty}(n - n_s)\theta p(n, 0)$ is the mean abandonment (reneging) rate in equilibrium and $E[N_{q_e}]$ is the mean number of customers in the system given by

$$E[N_{q_e}] = \sum_{n=0}^{n_s} np(n, 1) + \sum_{n=0}^{\infty} np(n, 0).$$

For this analysis, we consider three sets of numerical experiments. The first set studies the effect of the fraction of time that the system is observable in Section 5.1, whereas the second studies the effect of the duration of the unobservable periods in Section 5.2. Finally, the third studies the influence of the percentage of the entrance fee which is refundable when a customer reneges (i.e., the fraction r/f_e) in Section 5.3.

5.1 Fraction of time that the system is observable

First, we study the effect on strategic customer behavior of the fraction of time that the system is observable, $\gamma \in [0, 1]$, when the mean information cycle of the system – consisting of an unobservable and an observable period – is kept fixed. To this end, for a given mean information cycle of the system B, we set $\theta = \frac{1}{(1-\gamma)B}$ and $\zeta = \frac{1}{\gamma B}$, so that $1/\theta + 1/\zeta = B$, and let γ vary in $[0, 1]$. Our objective is to study whether there exists an ideal fraction of time, strictly between 0 and 1, for which the system should be observable given the duration B.

We first consider an instance with fixed mean information cycle, $B = 0.1$ (high frequency of alternations), and provide the plots of the joining probability q_e, the throughput μ_e and the social welfare S_e, when the customers follow the equilibrium strategy, as functions of γ in the three panels of Figure 3. In each panel, we plot three curves for the arrival rates $\lambda = 0.8, 1.1$ and 2.3, respectively. The rest of the parameters are kept fixed: the service rate is set $\mu = 1$ and the economic parameters are $R = 4$, $C = 1$, $f_e = f_s = 0$ and $r = -30$. The choices $B = 0.1$ and
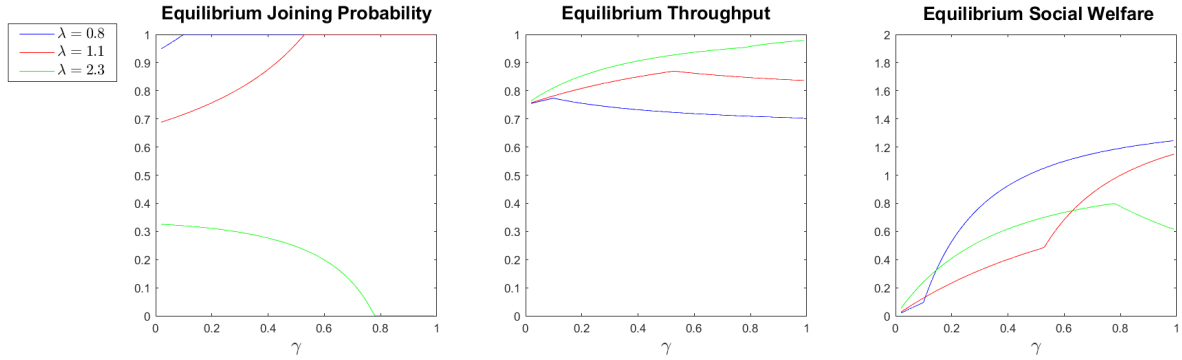

Figure 3: Customer's joining probability, throughput and social welfare with respect to γ for $\lambda = 0.8, 1.1, 2.3$ and for $B = 0.1, \mu = 1$, $R = 4$, $C = 1$, $f_e = f_s = 0$ and $r = -30$.

$r = -30$ represent that the alternations in the information periods are almost instantaneous and that reneging is prohibited. In this case, γ corresponds to the fraction of customers that observe the queue length upon arrival. The rest of the parameters are the same as in Hu, Li and Wang [15]. The resulting plots of Figure 3 coincide with their findings, cf. Figures 3, 4 and 5 in [15], and confirm that their model can be derived as a limiting case of the alternating information structure.

The present model allows a greater degree of control on the fraction of time that the system is observable and hence, it is not surprising that the equilibrium throughput is usually a unimodal function of γ. This is in agreement with the findings in [15], who proved that the equilibrium throughput and the equilibrium social welfare are in general unimodal and not monotonous in the fraction γ of informed customers. Similarly, Chen and Frank [3] have shown that regarding equilibrium throughput maximization, the observable version of the M/M/1 queue is preferable for high values of λ, whereas the unobservable version is preferable for low values of λ. Indeed, when λ is high, in the unobservable version no customer enters, whereas some customers do enter in the observable counterpart (those few that find the system in low congestion). The opposite happens when λ is low, i.e., all customers enter in the unobservable case, whereas not all customers enter in the observable case.

Regulating γ and keeping all other parameters fixed has multiple effects on the system. In case of low arrival rates λ, an increase in γ increases the customer entrance probability for the unobservable periods, but it also increases the abandonment probability since the system passes quickly from the unobservable to the observable mode. This double effect is reversed in the case of high arrival rates. Moreover, tuning γ changes the composition of the customers' population, which in turn, changes the join-or-balk game among the customers. The population consists of two different subpopulations of customers and the solution of the game becomes more intricate.

To understand the effect of γ for different frequencies of alternation between observable and unobservable periods, we perform a second experiment with increasing values of the mean information cycle $B = 0.1, 10, 100$. We set the arrival rate at $\lambda = 1.1$ (also used by [15]) and keep the rest of the parameters as in Figure 3. The results are shown in Figure 4. The main
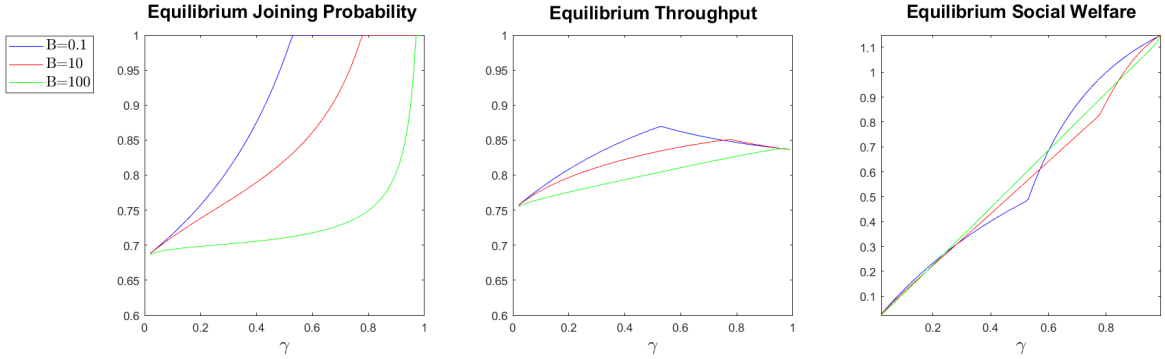

Figure 4: Customer's joining probability, throughput and social welfare with respect to γ for $B = 0.1$, 10, 100 and for $\lambda = 1.1$, $\mu = 1$, $R = 4$, $C = 1$, $f_e = f_s = 0$ and $r = -30$.

observation in Figure 4 is that given sufficient control on the duration of the information cycle, the alternating information structure can lead to increased social welfare in comparison to the information heterogeneity of [15]. This is achieved for intermediate values of γ for which the line for $B = 0.1$ ([15]) lies below the lines of both $B = 10$ and $B = 100$ (longer information cycles and hence, lower frequency of alternations). To allow comparisons, reneging has been kept prohibited in all these plots by setting $r = -30$. However, allowing customers to renege can further improve system performance.

A second feature that is revealed by the plots of Figures 3 and 4 is the concurrence of the value of γ in which the equilibrium joining probability hits 1, the mode in the equilibrium throughput and the tipping point in social welfare. This remains true for the social welfare if the joining probability hits instead 0, but not for the equilibrium throughput which remains increasing, cf. plots for $\lambda = 2.3$ in Figure 3 (the terms *increasing* and *decreasing* should be understood in a weak manner, i.e., non-decreasing and non-increasing, respectively). The explanation is based again on the previous discussion: as γ increases, it influences the joining and abandonment equilibrium probabilities towards the same direction (both probabilities increase or decrease). But when the joining probability hits its extreme value, 1 or 0, as γ varies, the effect of further increasing γ on the joining probability ceases to exist, whereas the effect on the abandonment probability continues. In short, Figures 3 and 4 illustrate the following

- The equilibrium joining probability is increasing in γ for low values of λ and decreasing in γ for high values of λ. The equilibrium throughput is an increasing or unimodal function of γ. Its mode coincides with the point at which q_e reaches 1. The equilibrium social welfare slope changes abruptly when q_e reaches 1. For certain values of γ, it is higher for information cycles of higher duration B.

- The equilibrium joining probability is decreasing in λ, the equilibrium throughput is increasing in λ, whereas the equilibrium social welfare is non-monotonic in λ.

18

- The equilibrium joining probability is decreasing in B, whereas the equilibrium throughput and the equilibrium social welfare are non-monotonic in B.

5.2 Duration of unobservable periods

Next, we turn to the effect of the duration of the unobservable periods, as expressed by varying values of θ, on strategic customer behavior. At the limiting case, in which the duration of the corresponding observable periods is very short (almost instantaneous), i.e., for large values of ζ, the alternating information structure reduces to an *announcement model* with announcement rate θ. This is precisely the setting studied in Burnetas, Economou and Vasiliadis [2]. In their framework, [2] show that the equilibrium throughput is a non-monotonic function of the announcement rate θ, when all other parameters are kept fixed.

To recover this result, in the first scenario, we let θ vary in $(0, 10)$ for three different arrival rates $\lambda = 7, 10, 40$ and select $\zeta = 300$ to model very short observable periods (almost instantaneous announcements). In all cases, the economic parameters $R = 5$, $C = 10$, and $f_e = f_s = r = 0$ and the operational parameter $\mu = 8$ have been kept fixed. The equilibrium perfomance measurers q_e, μ_e and S_e are plotted as functions of θ in the three panels of Figure 5. The three curves in each panel correspond to the different values of λ.
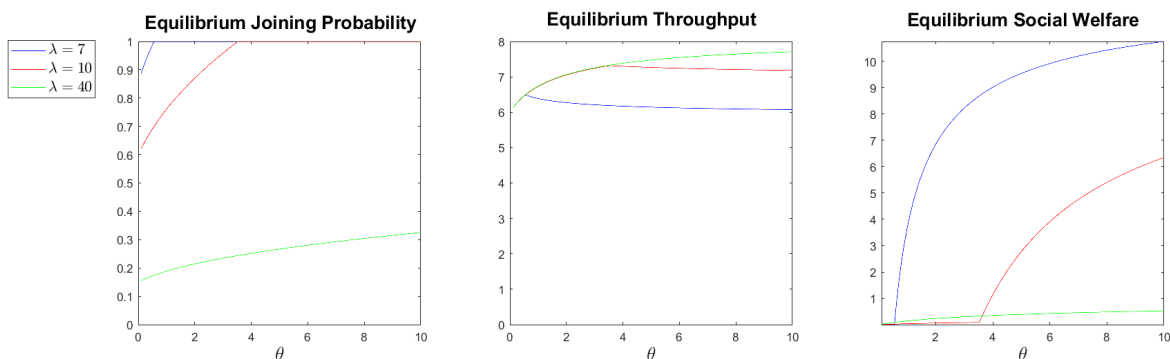

Figure 5: Equilibrium joining probability, throughput and social welfare with respect to θ, for $\lambda = 7, 10, 40$, when $\zeta = 300, \mu = 8$, $R = 5$, $C = 10$, $f_e = f_s = r = 0$.

The most interesting finding is that the ideal rate θ for maximizing equilibrium throughput lies strictly between 0 and ∞. This can be seen for the curves that correspond to $\lambda = 7$ and 10 and is also true for the curve with $\lambda = 40$ which attains its mode outside the selected range of θ. This happens because increasing θ has two opposing effects: On the one hand, it increases the equilibrium joining probability, because the uninformed customers become more willing to enter, knowing that they will be informed more quickly. On the other hand, it increases the reneging probability, because an uninformed customer who has joined the system during a high congestion period will abandon the system earlier. The trade-off between the two effects is not clear and this is the reason for the unimodality of the throughput.

The flexibility of the current information structure can be utilized to improve over the benchmark model of [2]. To see this, we consider a second experiment with decreased values of $\zeta = 1, 10, 100$. Lower values of ζ correspond to longer observable periods. We set the arrival rate at $\lambda = 40$, and keep the rest of the parameters as in Figure 5. We let θ to take values in $[0, 200]$. The results on the equilibrium performance measures q_e, μ_e and S_e are shown in the three panels of Figure 6.

The main conclusion that can be drawn from the diversity of the plots is the high dependence of the performance measures, in particular of the equilibrium joining probability and social welfare, on the interplay between the operational model parameters. Due to the their opposing effects, general statements cannot be formulated for broad ranges of parameters' values. For
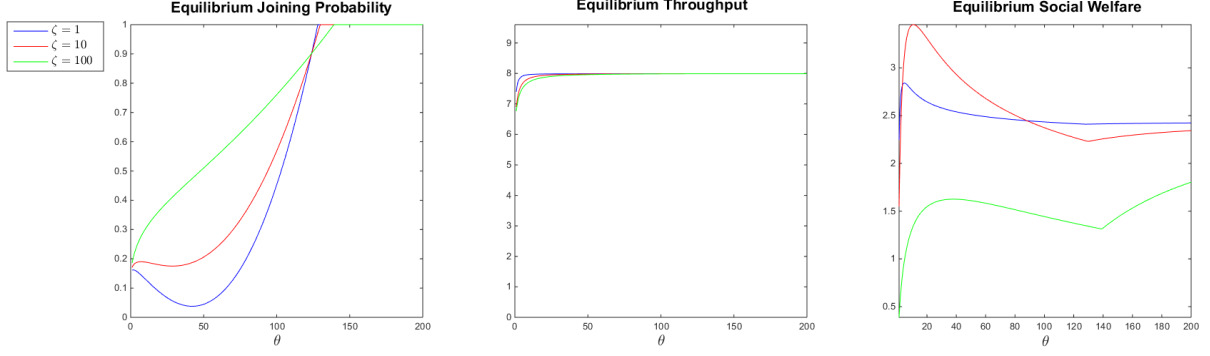
Figure 6: Customer's joining probability, throughput and social welfare with respect to θ for $\zeta = 1, 10, 100$ and for $\lambda = 40, \mu = 8$, $R = 5$, $C = 10$, $f_e = f_s = r = 0$.

practical situations, this suggests the necessity for a case-by-case analysis to test each set of candidate parameters individually but also indicates the broad applicability of the present model. From an analytic perspective, it underlines the significance to derive the equilibrium of the system and perform in turn comparative statics via the currently employed numerical methods and tools.

More concretely, as can be seen from the first panel in Figure 6, the equilibrium joining probability, is eventually increasing in θ and ultimately reaches 1 for any value of ζ. As θ increases, the uninformed customers know that they will learn the system state very quickly, so they have a strong incentive to join, independently of the value of ζ. If the observable periods are not long, here $\zeta = 100$, then uninformed customers have an incentive to join as θ increases, since reneging becomes easier. For longer observable periods, here $\zeta = 1$, the system becomes occupied by the informed customers which disincentivizes uninformed customers to join and explains the drop in the $\zeta = 1$ curve in the first panel for intermediate values of θ.

Finally, the equilibrium throughput quickly reaches its maximal possible value $\mu_e = 8$, whereas the social welfare behaves non-monotonically in both θ and ζ after an initial steep increase for low values of θ. Indeed, for values of θ close to 0, the model essentially corresponds to an unobservable queue. Since the arrival rate, $\lambda = 40$, is much higher than the service rate, $\mu = 1$, uninformed customers are incentivized to balk, (see also [3]). This leaves the system uncongensted for customer arriving at the short observable periods and leads to a sharp increase in equilibrium social welfare. However, as θ increases further, the effects become mixed. Again, an abrupt change in all curves occurs when the equilibrium joining probability hits 1. In short, some general statements that can be formulated based on the results in Figures 5 and 6 are the following

- The equilibrium joining probability is a non-monotonic function of θ. However, it is eventually increasing in θ and ultimately reaches 1. The equilibrium throughput is increasing or unimodal in θ, whereas the equilibrium social welfare is non-monotonic.
- The equilibrium joining probability is decreasing in λ, the equilibrium throughput is increasing in λ and the social welfare non-monotonic in λ.
- The equilibrium throughput is decreasing in ζ, whereas the equilibrium joining probability and social welfare are non-monotonic in ζ.

5.3 Fraction of refundable entrance fee

In the last set of experiments, we study the effect of the fraction of the entrance fee that is refundable, i.e., of r/f_e for $r \in [0, f_e]$, on the strategic customer behavior. The perfomance measures q_e, μ_e and S_e have been plotted in the three panels of Figure 7 as functions of r/f_e in $[0, 1]$ for three different service valuations, $R = 7, 10$ and 15. In all cases, the operational

parameters have been kept fixed at $\lambda = 1.3$, $\mu = 1$, $\zeta = 10$ and $\theta = 1$ and the remaining economic parameters at $C = 1$, $f_e = 5$ and $f_s = 0$.
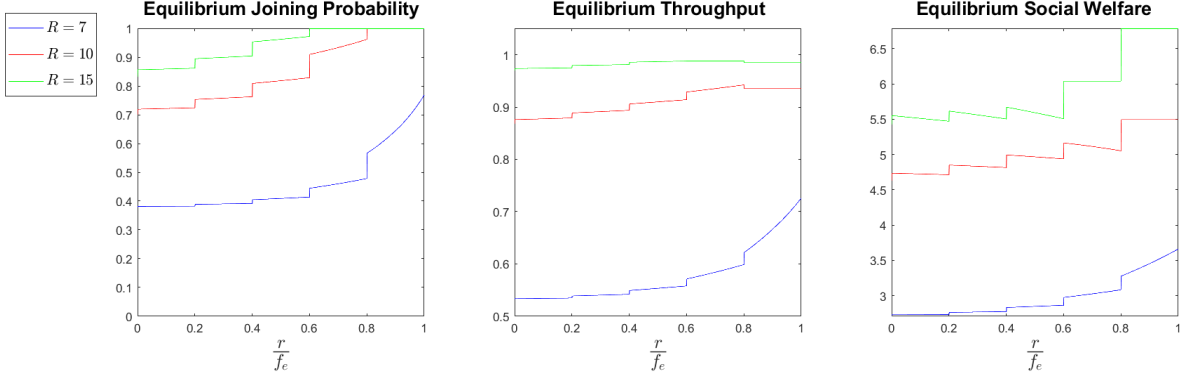

Figure 7: Customer's joining probability, throughput and social welfare with respect to $\frac{r}{f_e}$ for $R = 7$, 10, 15 and for $\lambda = 1.3$, $\mu = 1$, $\zeta = 10$, $\theta = 1$, $C = 1$, $f_e = 5$ and $f_s = 0$.

The main finding is the discontinuity of all equilibrium measures which is caused by the discrete changes in the reneging threshold n_s as r/f_e varies. At the points of change, the equilibrium measures undergo abrupt changes or jumps (depicted as vertical lines in the plots). In each interval of continuity, the value of n_s remains the same. Then, an increase in r/f_e makes the uninformed customers more willing to enter and abandon later if they find a high congestion. So, again the joining probability and the abandonment probability both increase and their trade-off is not clear.

In a second scenario, we use the same operational and economic parameters $\lambda = 1.3, \mu = 1, \zeta = 10, \theta = 1$, and $C = 1$, but we now fix $R = 7$ and $f_e + f_s = 5$. Again, we let r/f_e vary in $[0, 1]$. We examine three different scenarios for different values of $f_e = 1, 3$ and 5. Our aim is to model the situation in which the total fee is decomposed in two parts, the entrance and the service fees, and study effect of the percentage of the entrance fee that is refundable. The performance measures are plotted in the three panels of 8.
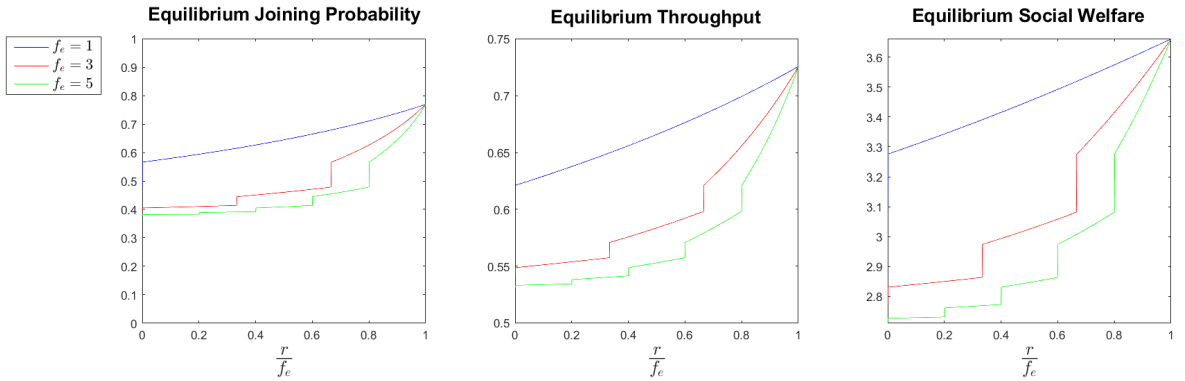

Figure 8: Customer's joining probability, throughput and social welfare with respect to $\frac{r}{f_e}$ for $f_e = 1$, 3, 5 and for $\lambda = 1.3$, $\mu = 1$, $\zeta = 10$, $\theta = 1$, $R = 7$, $C = 1$ and $f_s = 5 - f_e$.

The rest of the findings of Figures 7 and 8 are fairly intuitive. All equilibrium performance measures improve as the service valuation R increases or as the service fee f_s increases for a given sum $f_e + f_s$ of entrance and service fees. While the equilibrium joining probability and social welfare also improve as the ratio r/f_e tends to 1, this may not necessarily be true for the equilibrium throughput, which may decrease, at least marginally, as can be inferred from

the $R = 15$ and $R = 7$ plots in the middle panel of 7. Yet, in all other cases, the equilibrium increases as well. In short, the findings of Figures 7 and 8 are the following

- The equilibrium joining probability is an increasing function of r/f_e, the equilibrium throughput is an increasing or unimodal function of r/f_e and the equilibrium social welfare is monotonic in each interval of continuity but in general, it is a non-monotonic function of r/f_e. All equilibrium performance measures are discontinuous functions of r/f_e.
- All equilibrium performance measures are increasing in R and in the ratio f_s/f_e for constant $f_e + f_s$.

The results of all experiments are summarized in Table 2.

Effect	Variables		Equilibrium performance measures		
	axes	plots	q_e	μ_e	S_e
Fraction of time that the system is observable	γ		$\uparrow$ for low λ, $\downarrow$ for high λ	$\uparrow$ or $\wedge$	$\uparrow$ or $\wedge$
		λ	$\downarrow$	$\uparrow$	$\times$
		B	$\downarrow$	$\times$	$\times$
Duration of unobservable periods	θ		$\times$	$\uparrow$ or $\wedge$	$\times$
		λ	$\downarrow$	$\uparrow$	$\times$
		ζ	$\times$	$\downarrow$	$\times$
Fraction of refundable fee	r/f_e		$\uparrow$	$\uparrow$ or $\wedge$	$\times$
		R	$\uparrow$	$\uparrow$	$\uparrow$
		f_s/f_e	$\uparrow$	$\uparrow$	$\uparrow$

Table 2: Equilibrium performance measures in the numerical experiments: q_e denotes the equilibrium joining probability, μ_e the equilibrium throughput and S_e the equilibrium social welfare. Symbol $\uparrow$ stands for non-decreasing, $\downarrow$ for non-increasing, $\wedge$ for unimodal and $\times$ for non-monotonic nor unimodal. In column *Variables*, the field *axes* refers to the variables that appear in the horizontal axes of the panels in each figure and the field *plots* to the variables that yield the three different plots in each panel.

6 Summary and conclusions

In the present paper, we considered an M/M/1 queue that alternates between exponentially distributed observable and unobservable periods and which bridges the extremal cases of continutously observable and unobservable systems. While this model unifies and generalizes the existing approaches of [15] and [2], it remains analytically tractable since it always has a unique equilibrium that can be characterised via the system parameters. This allows for a comprehensive experimentation on the operational and economic system parameters to gain managerial insight. A main conclusion is that the equilibrium throughput and the corresponding social welfare are typically greater when an ideal level of alternation between observable and unobservable mode is used instead of the system being continuously observable or unobservable.

Our results imply that sufficient flexibility to control the information structure of a given queueing system improves its performance both from a managerial and a social perspective. Thus, apart from its practical relevance, the present model may also provide a benchmark for future studies in this direction. One interesting research problem is to extend the analysis in the case where the unobservable and observable periods are of constant lengths and not exponentially distributed. This seems quite difficult from an analytical point of view, but even

a numerical study deserves attention. Another interesting direction for future research concerns the case where the alternation between the observable and unobservable modes of the system is not static (i.e., specified by the exponential rates θ and ζ) as in the present study, but can be dynamically controlled by the administrator of the system.

References

[1] Artalejo, J. and Gómez-Corral (2008) *Retrial Queueing Systems, A Computational Approach.* Springer.

[2] Burnetas, A., Economou, A. and Vasiliadis, G. (2017) Strategic behavior in a queueing system with delayed observations. *Queueing Systems* **86**, 389-418.

[3] Chen, H. and Frank, M. (2004) Monopoly pricing when customers queue. *IIE Transactions* **36**, 569-581.

[4] Economou, A. and Grigoriou, M. (2015) Strategic balking behavior in a queueing system with a mixed observation structure. In *Proceedings of the 10th Conference on Stochastic Models of Manufacturing and Service Operations (SMMSO 2015), pp. 51-58*, University of Thessaly Press, Volos.

[5] Economou, A. and Kanta, S. (2008) Optimal balking strategies and pricing for the single server Markovian queue with compartmented waiting space. *Queueing Systems* **59**, 237-269.

[6] Edelson, N.M. and Hildebrand, K. (1975) Congestion tolls for Poisson queueing processes. *Econometrica* **43**, 81-92.

[7] Elaydi, S.N. (1996) *An Introduction to Difference Equations, 2nd Edition.* Springer, New York.

[8] Guo, P. and Zipkin, P. (2007) Analysis and comparison of queues with different levels of delay information. *Management Science* **53**, 962-970.

[9] Guo, P. and Zipkin, P. (2009) The effects of the availability of waiting-time information on a balking queue. *European Journal of Operational Research* **198**, 199-209.

[10] Hassin, R. (1986) Consumer information in markets with random products quality: The case of queues and balking. *Econometrica* **54**, 1185-1195.

[11] Hassin, R. (2016) *Rational Queueing.* CRC Press, Taylor and Francis Group, Boca Raton.

[12] Hassin, R. and Haviv, M. (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems.* Kluwer Academic Publishers, Boston.

[13] Hassin, R. and Koshman, A. (2014) Optimal control of a queue with high-low delay announcements: the significance of a queue. In *Valuetools Conference, 2014.*

[14] Hassin, R. and Roet-Green, R. (2014) The armchair decision: depart or stay home. *Working paper.*

[15] Hu, M., Li, Y. and Wang, J. (2018) Efficient ignorance: Information heterogeneity in a queue. *Management Science* **64**, 2650-2671.

[16] Ibrahim, R. (2018) Sharing delay information in service systems: A literature survey. *Queueing Systems* **89**, 49-79.

[17] Kim, B. and Kim, J. (2017) Optimal disclosure policies in a strategic queueing model. *Operations Research Letters* **45**, 181-186.

[18] Kulkarni, V.G. (2010) *Modeling and Analysis of Stochastic Systems, 2nd Edition.* CRC Press, Taylor and Francis Group, Boca Raton.

[19] Latouche, G. and Ramaswami, V. (1999) *Introduction to Matrix Analytic Methods in Stochastic Modeling.* ASA-SIAM series on Statistics and Applied Probability.

[20] Naor, P. (1969) The regulation of queue size by levying tolls. *Econometrica* **37**, 15-24.

[21] Roet-Green, R. (2013) *Information in queueing systems with strategic customers.* PhD thesis, School of Mathematical Sciences, Tel-Aviv Univesity.

[22] Shone, R., Knight, V.A. and Williams, J.E. (2013) Comparisons between observable and unobservable M/M/1 queues with respect to optimal customer behavior. *European Journal of Operational Research* **227**, 133-141.

[23] Simhon, E., Hayel, Y., Starobinski, D. and Zhu, Q. (2016) Optimal information disclosure policies in strategic queueing games. *Operations Research Letters* **44**, 109-113.

[24] Stidham, S. Jr. (2009) *Optimal Design of Queueing Systems.* CRC Press, Taylor and Francis Group, Boca Raton.

[25] Wang, J., Cui S. and Wang Z. (2018) Equilibrium Strategies in M/M/1 Priority Queues with Balking. *Production and Operations Management* **28**(1), 43–62.

Appendix A   Computation of the partial generating functions of the steady-state distribution

To obtain the partial generating functions, defined in (3.24) and (3.25), of the steady-state probabilities $p(n,i)$, when a given (n_e, n_s, q)-PES is used by the population of customers, we use the following approach. First, we multiply the balance equations for $p(n,i)$ in (3.6)-(3.13) by the appropriate power z^n and add them to obtain equations for the partial generating functions up to a few boundary probabilities to be determined later. We then solve the latter equations using standard algebraic methods.

A.1   Equations for the partial generating functions

By multiplying (3.7) with z^n and summing for $1 \leq n \leq n_e - 1$ we obtain

$$(\lambda + \mu + \zeta)P_{1a}(z) - \mu p(0,1) = \lambda \sum_{n=1}^{n_e-1} p(n-1,1)z^n + \theta \sum_{n=0}^{n_e-1} p(n,0)z^n + \mu \sum_{n=0}^{n_e-1} p(n+1,1)z^n$$

which reduces after straightforward algebraic manipulations to

$$\left[(\lambda + \mu + \zeta)z - \lambda z^2 - \mu\right] P_{1a}(z) - \theta z P_{0a}(z)$$
$$= \mu(z-1)p(0,1) - \lambda p(n_e-1,1)z^{n_e+1} + \mu p(n_e,1)z^{n_e}. \qquad (A.1)$$

Similarly, we derive a second equation for the generating functions $P_{0a}(z)$ and $P_{1a}(z)$, multiplying (3.12) by z^n and summing for $1 \leq n \leq n_s$, along with (3.11). A bit of algebra yields

$$\left[(\lambda q + \mu + \theta)z - \lambda q z^2 - \mu\right] P_{0a}(z) - \zeta z P_{1a}(z)$$
$$= \mu(z-1)p(0,0) - \lambda q p(n_e-1,0)z^{n_e+1} + \mu p(n_e,0)z^{n_e}. \qquad (A.2)$$

Equations (A.1), (A.2) can be written in matrix-form as

$$
\begin{bmatrix}
-\theta z & (\lambda + \mu + \zeta)z - \lambda z^2 - \mu \\
(\lambda q + \mu + \theta)z - \lambda q z^2 - \mu & -\zeta z
\end{bmatrix}
\cdot
\begin{bmatrix}
P_{0a}(z) \\
P_{1a}(z)
\end{bmatrix}
=
\begin{bmatrix}
N_{0a}(z) \\
N_{1a}(z)
\end{bmatrix}, \quad \text{(A.3)}
$$

where

$$
N_{0a}(z) = \mu(z-1)p(0,1) - \lambda p(n_e - 1, 1)z^{n_e+1} + \mu p(n_e, 1)z^{n_e} \quad \text{(A.4)}
$$

$$
N_{1a}(z) = \mu(z-1)p(0,0) - \lambda q p(n_e - 1, 0)z^{n_e+1} + \mu p(n_e, 0)z^{n_e}. \quad \text{(A.5)}
$$

Next, we derive a linear system for the generating functions $P_{0b}(z)$ and $P_{1b}(z)$ following the same procedure. More specifically, multiplying (3.9) with z^{n-n_e} for $n_e + 1 \leq n \leq n_s - 1$ and summing them together with (3.8) yields

$$
(\mu + \zeta) \sum_{n=n_e}^{n_s-1} p(n,1)z^{n-n_e} = \lambda p(n_e - 1, 1) + \theta \sum_{n=n_e}^{n_s-1} p(n,0)z^{n-n_e} + \mu \sum_{n=n_e}^{n_s-1} p(n+1,1)z^{n-n_e}.
$$

which can be written in a simplified form as

$$
[(\mu + \zeta)z - \mu] P_{1b}(z) - \theta z P_{0b}(z) = \lambda z p(n_e - 1, 1) - \mu p(n_e, 1) + \mu p(n_s, 1)z^{n_s-n_e}. \quad \text{(A.6)}
$$

A second equation for $P_{0b}(z)$ and $P_{1b}(z)$, can be derived from (3.12), multiplying with z^{n-n_e} and summing over all $n_e \leq n \leq n_s - 1$. It yields

$$
\begin{aligned}
&\left[(\lambda q + \mu + \theta)z - \lambda q z^2 - \mu \right] P_{0b}(z) - \zeta z P_{1b}(z) \\
&= -\lambda q p(n_s - 1, 0)z^{n_s-n_e+1} + \mu p(n_s, 0)z^{n_s-n_e} + \lambda q p(n_e - 1, 0)z - \mu p(n_e, 0). \quad \text{(A.7)}
\end{aligned}
$$

Again, (A.6), (A.7), can be written in matrix-form as

$$
\begin{bmatrix}
-\theta z & (\mu + \zeta)z - \mu \\
(\lambda q + \mu + \theta)z - \lambda q z^2 - \mu & -\zeta z
\end{bmatrix}
\cdot
\begin{bmatrix}
P_{0b}(z) \\
P_{1b}(z)
\end{bmatrix}
=
\begin{bmatrix}
N_{0b}(z) \\
N_{1b}(z)
\end{bmatrix}, \quad \text{(A.8)}
$$

where

$$
N_{0b}(z) = \mu p(n_s, 1)z^{n_s-n_e} + \lambda p(n_e - 1, 1)z - \mu p(n_e, 1), \quad \text{(A.9)}
$$

$$
N_{1b}(z) = -\lambda q p(n_s - 1, 0)z^{n_s-n_e+1} + \mu p(n_s, 0)z^{n_s-n_e} + \lambda q p(n_e - 1, 0)z - \mu p(n_e, 0). \quad \text{(A.10)}
$$

For deriving an equation for $P_{0c}(z)$, we multiply (3.12) and (3.13) with z^{n-n_s} and sum over all

$n \geq n_s$. We have that

$$(\lambda q + \mu + \theta) \sum_{n=n_s}^{\infty} p(n,0)z^{n-n_s} = \lambda q \sum_{n=n_s}^{\infty} p(n-1,0)z^{n-n_s} + \zeta p(n_s,1) + \mu \sum_{n=n_s}^{\infty} p(n+1,0)z^{n-n_s},$$

which reduces easily to

$$\left[(\lambda q + \mu + \theta)z - \lambda q z^2 - \mu \right] P_{0c}(z) = N_{0c}(z), \tag{A.11}$$

where $N_{0c}(z) = \lambda q p(n_s - 1, 0)z + \zeta p(n_s, 1)z - \mu p(n_s, 0)$. Hence, the balance equations (3.6)-(3.13) have been transformed into equations (A.3), (A.8) and (A.11) for the partial generating functions, which can be in turn easily expressed in closed form as rational functions of z via Cramer's rule. It remains to obtain the boundary probabilities that appear in $N_{0a}(z)$, $N_{0b}(z)$, $N_{0c}(z)$, $N_{1a}(z)$ and $N_{1b}(z)$. To this end, we will use the balance equation (3.10) and the normalization equation (3.14) that have not been used yet. However, these are only 2 equations in the 9 unknown boundary probabilities. The additional required equations will be derived from the roots of the determinants of the linear systems (A.3), (A.8) and the roots of the coefficient of $P_{0c}(z)$ in (A.11).

A.2 Roots of the denominators of the partial generating functions

Starting with the $P_{0c}(z)$, we define

$$D_c(z) = (\lambda q + \mu + \theta)z - \lambda q z^2 - \mu, \tag{A.12}$$

which is the coefficient of $P_{0c}(z)$ in (A.11). Since $D_c(0) = -\mu < 0$, $D_c(1) = \theta > 0$ and $\lim_{z \to \infty} D_c(z) = -\infty$, it follows from Bolzano's Theorem that there exist real roots $z_{c,1} \in (0,1)$ and $z_{c,2} \in (1,\infty)$ of $D_c(z)$ which are given by

$$z_{c,1}, z_{c,2} = \frac{\lambda q + \mu + \theta \mp \sqrt{(\lambda q + \mu + \theta)^2 - 4\lambda q \mu}}{2\lambda q}. \tag{A.13}$$

Next, we derive the roots of the determinant $D_b(z)$ of the linear system (A.8). We have that

$$D_b(z) = \det \begin{bmatrix} -\theta z & (\mu + \zeta)z - \mu \\ (\lambda q + \mu + \theta)z - \lambda q z^2 - \mu & -\zeta z \end{bmatrix}$$

$$= (\lambda q(\mu + \zeta)z^2 - (\lambda q + \mu + \theta + \zeta)\mu z + \mu^2)(z - 1).$$

Therefore, $D_b(z) = 0$ has three roots, i.e., $z_{b,1} = 1$ and

$$z_{b,2}, z_{b,3} = \frac{(\lambda q + \mu + \theta + \zeta)\mu \mp \sqrt{(\lambda q + \mu + \theta + \zeta)^2 \mu^2 - 4\lambda q(\mu + \zeta)\mu^2}}{2\lambda q(\mu + \zeta)}.$$

Similarly, for the determinant $D_a(z)$ of the linear system (A.3) we have

$$D_a(z) = \det \begin{bmatrix} -\theta z & (\lambda + \mu + \zeta)z - \lambda z^2 - \mu \\ (\lambda q + \mu + \theta)z - \lambda q z^2 - \mu & -\zeta z \end{bmatrix}$$

$$= (-\lambda^2 q z^3 + \lambda(\mu + \theta + (\lambda + \mu + \zeta)q)z^2 - \mu(\lambda q + \mu + \theta + \zeta + \lambda)z + \mu^2) \cdot (z - 1).$$

Therefore, $D_a(z) = 0$ has four roots, i.e., $z_{a,1} = 1$ and the three roots of the cubic equation

$$-\lambda^2 q z^3 + \lambda(\mu + \theta + (\lambda + \mu + \zeta)q)z^2 - \mu(\lambda q + \mu + \theta + \zeta + \lambda)z + \mu^2 = 0,$$

which can be calculated by the general formula for the roots of a cubic equation and are denoted as $z_{a,k}$, for $k = 2, 3, 4$.

A.3 Computation of the partial generating functions

In this section, we provide a simple procedure for the computation of the partial generating functions. Solving (A.11) for $P_{0c}(z)$ yields

$$P_{0c}(z) = \frac{(\lambda q p(n_s - 1, 0) + \zeta p(n_s, 1))z - \mu p(n_s, 0)}{[(\lambda q + \mu + \theta)z - \lambda q z^2 - \mu]}. \tag{A.14}$$

Since $z_{c,1}$ given by (A.13) (with the minus sign), is a root of the denominator of $P_{0c}(z)$ inside the closed unit disc, then it should necessarily be a root of its numerator, as $P_{0c}(z)$ is known to converge in the closed unit disc (as a probability generating function). Hence, the numerator in (A.14) is a multiple of $z - z_{c,1}$, whereas the denominator can be factored as $-\lambda q(z - z_{c,1})(z - z_{c,2})$. Thus, (A.14) can be rewritten as

$$P_{0c}(z) = \frac{C(z - z_{c,1})}{(z - z_{c,1})(z - z_{c,2})} = \frac{C}{z - z_{c,2}},$$

where C is a constant. But $P_{0c}(0) = p(n_s, 0)$, so we conclude that $C = -z_{c,2}\, p(n_s, 0)$. Recall, now, that $z_{c,1}, z_{c,2}$ are roots of the quadratic equation in (A.12), thus $z_{c,1} \cdot z_{c,2} = \frac{\mu}{\lambda q}$. Therefore $P_{0c}(z)$ assumes the form

$$P_{0c}(z) = \frac{p(n_s, 0)}{1 - \frac{\lambda q z_{c,1}}{\mu} z} = \sum_{n=n_s}^{\infty} p(n_s, 0)(\frac{\lambda q z_{c,1}}{\mu})^{n-n_s} z^{n-n_s}. \tag{A.15}$$

For the derivation of $P_{0b}(z), P_{1b}(z)$ and $P_{0a}(z), P_{1a}(z)$, we apply Cramer's rule to the linear systems (A.8) and (A.3), respectively. Therefore, for $z \neq z_{1,b}, z_{2,b}, z_{3,b}$, we have

$$
\begin{bmatrix} P_{0b}(z) \\ P_{1b}(z) \end{bmatrix} = \begin{bmatrix} \frac{-\zeta z N_{0b}(z) - [(\mu+\zeta)z - \mu] N_{1b}(z)}{D_b(z)} \\ \frac{-\theta z N_{1b}(z) - [(\lambda q + \mu + \theta)z - \lambda q z^2 - \mu] N_{0b}(z)}{D_b(z)} \end{bmatrix}, \tag{A.16}
$$

and for $z \neq z_{1,a}, z_{2,a}, z_{3,a}, z_{4,a}$ we have

$$
\begin{bmatrix} P_{0a}(z) \\ P_{1a}(z) \end{bmatrix} = \begin{bmatrix} \frac{-\zeta z N_{0a}(z) - [(\lambda+\mu+\zeta)z - \lambda z^2 - \mu] N_{1a}(z)}{D_a(z)} \\ \frac{-\theta z N_{1a}(z) - [(\lambda q + \mu + \theta)z - \lambda q z^2 - \mu] N_{0a}(z)}{D_a(z)} \end{bmatrix}. \tag{A.17}
$$

Now, we have to compute the boundary probabilities that appear in $N_{0a}(z)$, $N_{0b}(z)$, $N_{1a}(z)$, $N_{1b}(z)$ and $p(n_s, 0)$. These are 9 probabilities: $p(0,0)$, $p(0,1)$, $p(n_e-1,0)$, $p(n_e-1,1)$, $p(n_e,0)$, $p(n_e,1)$, $p(n_s-1,0)$, $p(n_s,0)$ and $p(n_s,1)$ (see (A.4),(A.5),(A.9),(A.10) and (A.15)).

Rewriting (3.10) (that has not been used for the derivation of the equations that govern the partial generating functions) in terms of $P_{0c}(z)$ yields

$$
(\mu + \zeta)p(n_s, 1) = \theta P_{0c}(1) = \frac{\theta p(n_s, 0)}{1 - \frac{\lambda q z_{c,1}}{\mu}}
$$

and, we obtain

$$
p(n_s, 1) = \frac{\theta}{(\mu + \zeta)(1 - \frac{\lambda q z_{c,1}}{\mu})} p(n_s, 0). \tag{A.18}
$$

So, we have expressed $p(n_s, 1)$ in terms of $p(n_s, 0)$. Next, to obtain the rest 7 unknown probabilities in terms of $p(n_s, 0)$, we exploit the fact that the numerator of $P_{0b}(z)$ given in (A.16), and the numerator of $P_{0a}(z)$ given in (A.17) should vanish for $z = z_{b,1}, z_{b,2}, z_{b,3}$ and for $z = z_{a,1}, z_{a,2}, z_{a,3}, z_{a,4}$, respectively, because these partial generating functions are polynomials and cannot have singularities (poles). Therefore, we obtain the following 7 equations, one for each root of the corresponding denominator, to obtain the remaining 7 probabilities $p(0,0)$, $p(0,1)$, $p(n_e-1,0)$, $p(n_e-1,1)$, $p(n_e,0)$, $p(n_e,1)$ and $p(n_s-1,1)$ in terms of $p(n_s, 0)$. These are:

$$
-\zeta z_{b,i} N_{0b}(z_{b,i}) - [(\mu + \zeta)z_{b,i} - \mu] N_{1b}(z_{b,i}) = 0, \qquad \text{for } i = 1, 2, 3, \tag{A.19}
$$

$$
-\zeta z_{a,i} N_{0a}(z_{a,i}) - [(\lambda + \mu + \zeta)z_{a,i} - \lambda z_{a,i}^2 - \mu] N_{1a}(z_{a,i}) = 0, \qquad \text{for } i = 1, 2, 3, 4. \tag{A.20}
$$

Finally, $p(n_s, 0)$ is determined using the normalizing equation, and the derivation of the partial generating functions is completed. In practice, one assigns an arbitrary positive value to $p(n_s, 0)$ (e.g., $p(n_s, 0) = 1$), then computes the other boundary probabilities using (A.18) and solving the linear system of (A.19) and (A.20) and finally normalizes the solution so that the total steady-state probability be 1.