https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 69, No. 7, July 2023, pp. 3778–3796
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Information Design for Congested Social Services: Optimal Need-Based Persuasion

Jerry Anunrojwong,[a] Krishnamurthy Iyer,[b,*] Vahideh Manshadi[c]

[a] Columbia Business School, Columbia University, New York, New York 10027; [b] Industrial and Systems Engineering, University of Minnesota, Minneapolis, Minnesota 55455; [c] Yale School of Management, Yale University, New Haven, Connecticut 06511
*Corresponding author

Contact: jerryanunroj@gmail.com, https://orcid.org/0000-0001-8422-9539 (JA); kriyer@umn.edu, https://orcid.org/0000-0002-5538-1432 (KI); vahideh.manshadi@yale.edu, https://orcid.org/0000-0001-9103-7797 (VM)

Received: August 13, 2020
Revised: May 19, 2021; October 29, 2021
Accepted: January 12, 2022
Published Online in Articles in Advance: October 14, 2022

https://doi.org/10.1287/mnsc.2022.4548

Abstract. We study the effectiveness of information design in reducing congestion in social services catering to users with varied levels of need. In the absence of price discrimination and centralized admission, the provider relies on sharing information about wait times to improve welfare. We consider a stylized model with heterogeneous users who differ in their private outside options: *low-need* users have an acceptable outside option to the social service, whereas *high-need* users have no viable outside option. Upon arrival, a user decides to wait for the service by joining an unobservable first-come-first-serve queue, or leave and seek her outside option. To reduce congestion and improve social outcomes, the service provider seeks to persuade more low-need users to avail their outside option, and thus better serve high-need users. We characterize the Pareto-efficient signaling mechanisms and compare their welfare outcomes against several benchmarks. We show that if either type is the overwhelming majority of the population, then information design does not provide improvement over sharing full information or no information. On the other hand, when the population is sufficiently heterogeneous, information design not only Pareto-dominates full-information and no-information mechanisms, in some regimes it also achieves the same welfare as the "first-best," that is, the Pareto-efficient centralized admission policy with knowledge of users' types.

History: Accepted by Gabriel Weintraub, revenue management and market analytics.
Funding: This work was supported by the National Science Foundation, Division of Civil, Mechanical and Manufacturing Innovation [Grants CMMI-2002155 and CMMI-2002156].
Supplemental Material: The data and e-companion are available at https://doi.org/10.1287/mnsc.2022.4548.

Keywords: information design • social services • Pareto improvement • congestion

1. Introduction

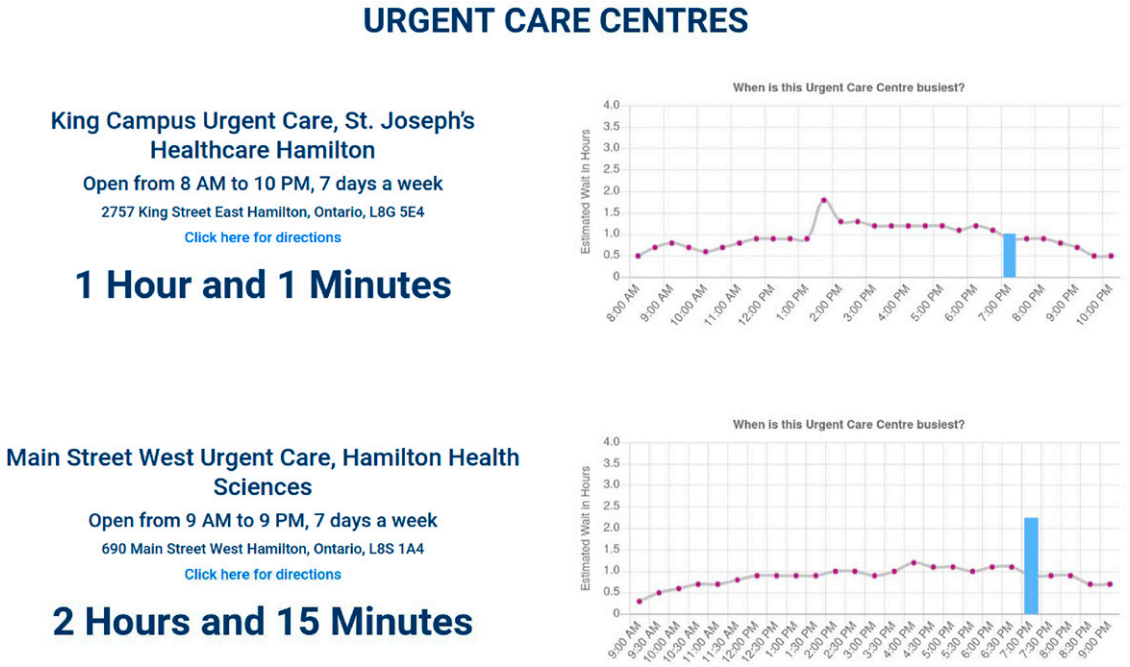
Social services often face the challenge of congestion due to their limited capacity relative to their demand. The congestion partly stems from the inclusionary intent of such services: a toll-free road is available to all citizens, even those who can afford alternative tolled ones; a broad range of low- and middle-income households are eligible to apply for public housing; urgent care centers admit patients with varied levels of condition severity. How can a social service provider reduce congestion and thus the efficiency loss associated with service delay?

In this context, the two controls commonly used for managing congestion, namely, pricing and centralized admission control, are inapplicable due to fairness and implementation considerations. However, the service provider may have control over information about the status of the system to be shared with users. As such, the service provider can leverage this informational advantage to influence the consumer's decision in seeking the social service.

1.1. Motivating Examples

A wide range of information provision policies are employed in practice. In the context of urgent care, some hospitals aim to provide real-time estimates of wait times to patients. For example, see Figure 1 for a snapshot of the Hamilton Healthcare System's wait-time dashboard, which we discuss further below (see also JFK Medical Center and San Mateo Medical Center, which employ similar programs).[1] On the other hand, in the context of public housing, certain authorities provide no wait-time information (see, e.g., Housing Authority of the County of Alameda), whereas others provide average estimates (see, e.g., New York Public Housing and Project-Based Voucher Waiting Lists). We highlight that in the aforementioned applications—which we broadly refer to as *social services*—managing congestion by

Figure 1. (Color online) Screenshot of the Wait-Time Dashboard Program for the Hamilton Healthcare System



"pricing out" users or by controlling admissions is impractical or undesirable.

Through information provision, service providers aim to not only inform users about their wait times but also to "help" users decide whether to seek the service. We take the Hamilton Healthcare System as our leading example. As reported in Mitchell (2020), upon launching their wait-time dashboard program, managers envision that providing wait-time information is particularly useful for patients with *less severe* conditions who can use this information to decide whether to currently seek care at a particular emergency center. Here, we highlight a quote from one manager:[2]

> There are still [going to] be people who have services like nephrology, or their heart doctors, or their lung doctors, who should go regardless of the wait time. . . . But for those that have less serious conditions, they can decide not only where but when to go.

The insights mentioned above highlight that, in applications such as emergency care, there are fundamental differences in the level of need in the user population: some have no choice but to seek the service regardless of the congestion level, whereas others can forgo the service if they perceive that the wait time is too long.

It is this fundamental heterogeneity of need that healthcare systems rely on when using wait-time dashboard programs, like that of the Hamilton healthcare system, to manage congestion. In this context, rather than providing full information, one can design dashboards that provide *coarsened* information about the congestion. For instance, a dashboard may announce that the wait time is above or below a threshold x, or in between a sequence of thresholds $x_1, x_2, \ldots, x_k$. Such coarsened information induce a belief that the congestion level might be high even in times of moderate congestion, and thus could persuade away users with less severe need from seeking service, resulting in reduced congestion overall. In this paper, we study the effectiveness of such information provision policies.

1.2. Overview of Our Work

To investigate the effectiveness of information design in improving welfare for a congested social service, we develop a stylized model that captures the key features of such a system. We consider a single-server queueing system where users arrive according to a Poisson process and their service times are independent and identically distributed and exponentially distributed.[3] Upon arrival, each user decides to either wait for the service by joining an unobservable queue or seek her outside option. To capture the disparity that users face with regard to the quality of their outside options, we categorize users into two groups: (1) *high-need* users who have no feasible outside option and (2) *low-need* users who have a viable alternative. Both types incur higher waiting costs upon joining a longer queue. Upon arrival, a high-need user always joins, as she does not

have any other choice. However, a low-need user makes a *join* or *leave* decision to maximize her expected utility. Even though an arriving user does not observe the queue, her decision relies on her belief about the queue size based on the information shared by the service provider.

We assume that the service provider has complete information about the status of the queue and that he can decide how much of this information he will share with the arriving user. Sharing the information fully may lead to bad welfare outcomes because a utility-maximizing user does not internalize the negative externality that she imposes on others (Naor 1969). Instead, the service provider can use the lever of information sharing to influence users' beliefs about the queue size and, consequently, their decisions. We adopt the framework of Bayesian persuasion or information design[4] (Kamenica and Gentzkow 2011) in which the service provider commits[5] to a signaling mechanism in response to which users follow an equilibrium strategy. The welfare of each type is thus determined by the signaling mechanism and the corresponding equilibrium response of the users. Because high-need users always join the queue, the service provider does not need to know user types to implement a signaling mechanism.

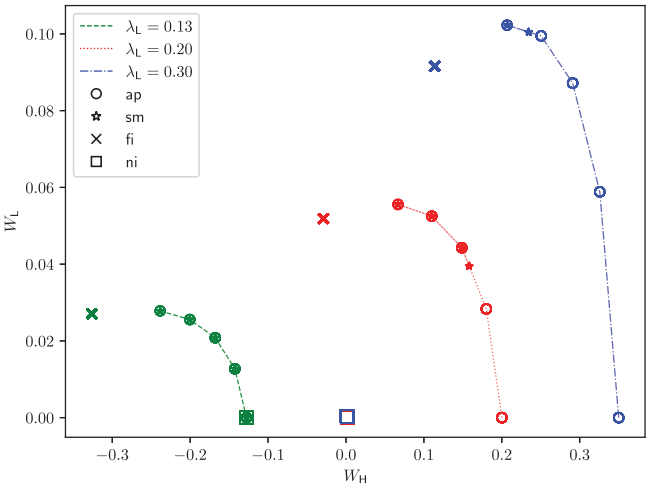
Our analysis follows the standard approach (see, e.g., Bergemann and Morris 2016, Candogan and Drakopoulos 2019, Lingenbrink and Iyer 2019) which allows us to only consider *obedient* binary signaling mechanisms where, upon the arrival of a user, the service provider makes a "join" or "leave" recommendation and the user finds it incentive-compatible to follow that recommendation. Further, it builds on Lingenbrink and Iyer (2019) to establish an equivalence between the class of obedient binary signaling mechanisms and the set of steady-state distributions that satisfy certain linear constraints. To ensure welfare improvement for *both* types, we focus on Pareto-efficient signaling mechanisms and establish structural results for any such mechanisms. Under mild monotonicity assumptions on utility functions, we show that any Pareto-efficient signaling mechanism has a threshold structure (Theorem 1).

With these structural results, we compare the optimal signaling mechanism against the benchmarks of full-information sharing and no-information sharing. Our analysis reveals several intriguing insights into effectiveness of information design. First, there exists a signaling mechanism that Pareto-dominates full-information sharing unless the latter remains Pareto-efficient even if the service provider is allowed to disregard user incentives (Proposition 3 and Theorem 4). However, if the population is mostly comprised of low-need users, then the welfare gain due to information design is fairly limited (Proposition 1). This dichotomy stems from the intuition that in the absence of high-need users, a low-need user cannot be persuaded to leave if the queue length falls below the threshold up to which she would have joined under full information. On the other hand, when high-need users are present, it is possible to persuade more low-need users to leave. Second, there exists a signaling mechanism that Pareto-dominates no information sharing only if the arrival rate of high-need users does not exceed a threshold. However, if high-need users constitute the overwhelming majority, then, interestingly, no information is Pareto-efficient, even when the provider is allowed to disregard user incentives (Proposition 3 and Theorem 5). The main intuition behind this result is that, with abundance of demand from high-need users, the system is so congested that a low-need user does not need much persuasion to choose her outside option over the social service. Conversely, if the system is not overcrowded by high-need users, then information design proves effective over sharing no information. Putting these insights together, *we conclude that signaling is particularly effective if the user population shows sufficient heterogeneity in need.*

To further study the power of information design, we compare its Pareto frontier with that of a strong benchmark in which the service provider implements a Pareto-efficient *admission policy* disregarding the user's incentives. Interestingly, we show that if the arrival rate of high-need users is higher than a threshold, then the two Pareto frontiers indeed coincide. Even if the arrival rate of high-need users is below that threshold, the two Pareto frontiers show considerable overlap (Theorem 6). This further illustrates the effectiveness of information design: any Pareto-efficient signaling mechanism that belongs to the overlapping regions of the frontier achieves the same welfare outcomes as those of an admission policy that cannot only observe the user types, but also enforces the join or leave decision without regard to their incentives. Further, in such cases, no user is indifferent between their recommended action and the alternative, implying that the signaling mechanism primarily plays the role of a coordination device. This is in contrast with usual persuasion settings, where the optimal signaling mechanism extracts all user surplus for some signals.

To highlight our comparative insights, in Section 6, we complement our theoretical findings with illustrative numerical examples (see Figures 2–4 and their related discussions). Additionally, in Section 7, we describe how our model can be generalized to incorporate a finite outside option for high-need users (in Section 7.1) and heterogeneity in service rates (in Section 7.2).[6] We analytically or numerically show our qualitative insights hold in these richer models (see Proposition 4, Figures 5–6, and their related discussions).

Figure 2. (Color online) Welfare of Pareto-Efficient Signaling Mechanisms and Admission Policies for $\lambda_L \in \{0.13, 0.20, 0.30\}$, $\lambda_H = 1 - \lambda_L$, and $c = 0.15$



Notes. Here, green (dashes) represents $\lambda_L = 0.13$, red (dots) represents $\lambda_L = 0.20$, and blue (dash-dots) represents $\lambda_L = 0.30$. Further, circles (○) represent efficient admission policies (ap), stars (⋆) represent efficient signaling mechanisms (sm), the cross (×) represents the full-information mechanism (fi), and the square (□) represents the no-information mechanism (ni). (The no-information points for $\lambda_L \in \{0.2, 0.3\}$ overlap, and so do those corresponding to signaling mechanisms and admission policies for each fixed λ_L.)

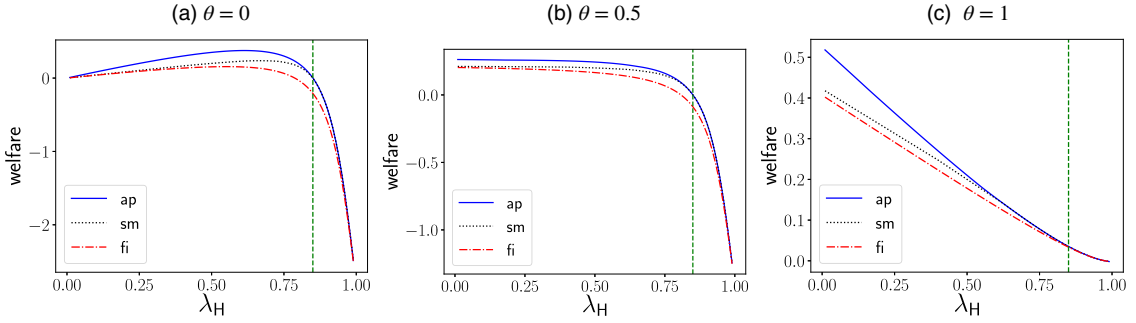
1.3. Managerial Insights

In summary, our work investigates the effectiveness of information design as a potential approach for reducing congestion in social services offered to users heterogeneous in their needs. Using a stylized model, we show that by implementing a Pareto-efficient signaling mechanism, the service provider can achieve Pareto improvement in the welfare by persuading more low-need users to seek an outside option, thereby reducing congestion. We also identify conditions under which information design not only outperforms the simple mechanisms of full- or no-information sharing but also achieves the same welfare outcomes as centralized admission policies that know each user's need for the service.

As wait-time dashboard programs have become prevalent means for congestion management in service systems, there is a natural impulse to design systems that accurately estimate and share complete information. However, contrary to general wisdom, our results show that sharing accurate information could in fact be uniformly detrimental to all the users. Instead, revealing partial information, say, in the form of thresholds and/or intervals, can improve the welfare outcomes across all users. Dashboard programs based on such coarsened information would also be practically appealing, as they alleviate the need to accurately estimate wait times in real time, a task that has been documented to be significantly challenging in practice (Ang et al. 2016, Xavier 2017). Thus, our results imply that information design not only alleviates the need for accurate wait-time estimation, this benefit also comes at no welfare cost.

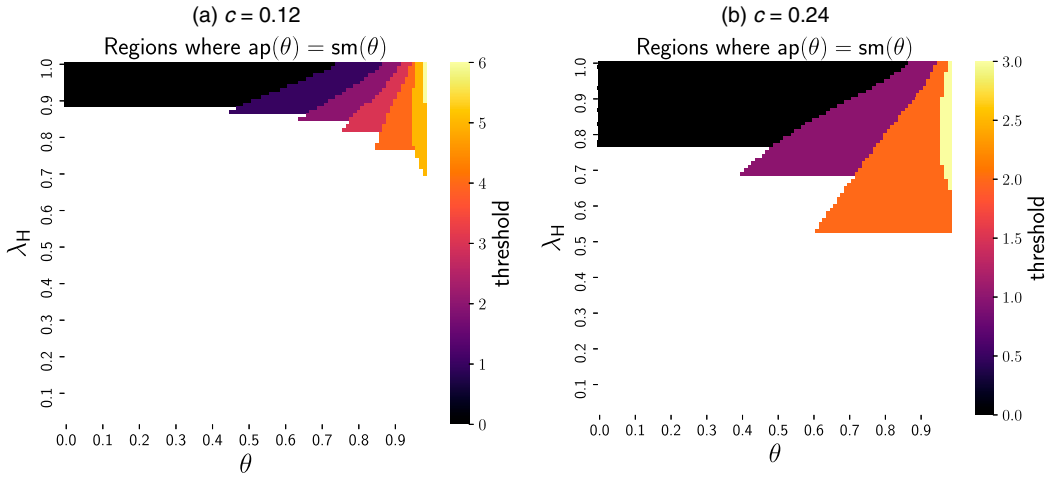
Lastly, we discuss two practical concerns that one may have about disclosing partial information for social services: repugnance and information leakage. With regard to the former, given that most information provision policies commonly used in practice do not follow full-information disclosure, we do not envision that implementing our proposed policies would be perceived differently.[7] With regard to the latter, we emphasize that we only focus on public signaling mechanisms, which removes the possibility of information leakage across agents. However, information leakage over time can happen: if an agent strategically waits upon arrival and observes more than one signal before deciding to join or leave, then she may be able to infer the state of the system more accurately. Nevertheless, our prescribed policy of only disclosing whether the congestion level is above/below a threshold would still perform well: observing a few signals only reveals extra information if the congestion level is close to the threshold and, thus, the signal changes. Consequently,

Figure 3. (Color online) Welfare of the Pareto-Efficient Signaling Mechanism sm(θ), the Pareto-Efficient Admission Policy ap(θ), and the Full-Information Mechanism fi for $\theta \in \{0, 0.5, 1\}$



Note. Here, $\lambda_L = 1 - \lambda_H$ and $c = 0.15$.

Figure 4. (Color online) Regions of the (θ, λ_H) Plane Where $\mathsf{ap}(\theta) = \mathsf{sm}(\theta)$, i.e., Where the Signaling Mechanism $\mathsf{sm}(\theta)$ is Pareto-Efficient Within Π_{AP}



Note. The colors represent the value of the threshold in $\mathsf{ap}(\theta)$.

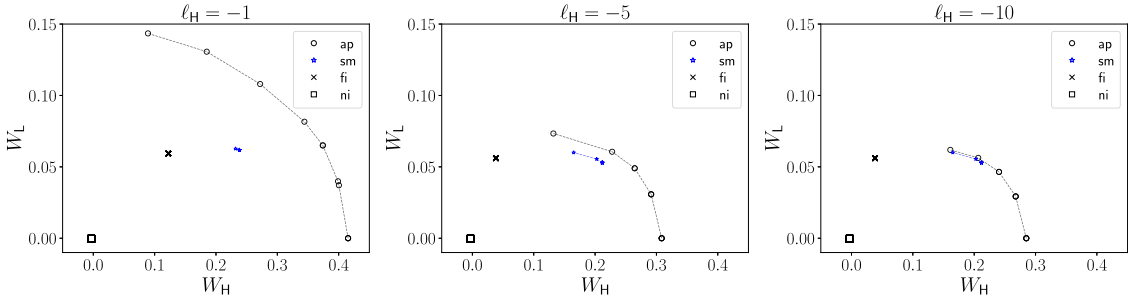
our policy still persuades away those who arrive when the congestion level is sufficiently above the threshold.[8]

1.4. Related Work

Our work relates to and contributes to several streams of literature.

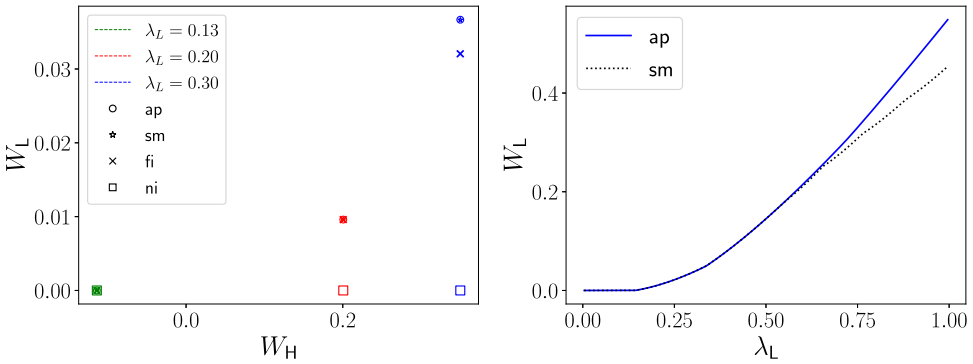
1.4.1. Information Design. Like ours, in many other settings service providers and platforms have access to more information than their customers. As such, informational aspects of service and platform operations have been studied in many applications. Adopting the framework of Bayesian persuasion pioneered by Kamenica and Gentzkow (2011), Lingenbrink and Iyer (2018) and Drakopoulos et al. (2018) study the effectiveness of information design for influencing the customers' time of purchase in order to maximize the platform's revenue. In a similar context, Küçükgül et al. (2019) study information design for time-locked sales campaigns on online platforms. Focusing on two-sided platforms, Bimpikis et al. (2020) examine the impact of information design on supply-side decisions toward the goal of increasing a platform's revenue. Kremer et al. (2014) and Papanastasiou et al. (2018) focus on information design in a sequential learning setting with the goal of maximizing social welfare. In the context of misinformation on social platforms, Candogan and Drakopoulos (2019) study how the platform can optimally signal the content accuracy while incentivizing desirable levels of user engagement in the presence of positive network externalities. Outside the framework of Bayesian persuasion, for dynamic contests, Bimpikis et al. (2019) show that the information disclosure policy used to inform participants about the status of competition substantially impacts the outcome. Kanoria and Saban (2021) show that a two-sided matching platform can significantly improve welfare by hiding information about the quality of a user's potential partners. In another interesting direction, Nahum et al. (2015) show that in two-sided matching, the presence of experts who can reveal information

Figure 5. Welfare of Pareto-Efficient Signaling Mechanisms and Admission Policies for $\lambda_L = 0.2$, $\lambda_H = 1 - \lambda_L$, $c = 0.13$, and $\ell_H = -1$ (Left Panel), $\ell_H = -5$ (Middle Panel), and $\ell_H = -10$ (Right Panel)



Note. In all three panels, circles (∘) represent efficient admission policies (ap), stars (⋆) represent efficient signaling mechanisms (sm), the cross (×) represents the full-information mechanism (fi), and the square (□) represents the no-information mechanism (ni).

Figure 6. (Color online) Welfare of Pareto-Efficient Signaling Mechanisms (Stars) and Admission Policies (Circles) and Welfare of L-Type Users Under the Pareto-Efficient Signaling Mechanism and Admission Policy



Notes. Left: Welfare of Pareto-efficient signaling mechanisms and admission policies for a preemptive priority queue with $\mu_L = 1.05$, $\mu_H = 1$, and $c_L = c_H = 0.15$. The colors and shapes have the same definition as in Figure 2. Right: The welfare of L-type users under sm and ap for $\lambda_L \in [0,1]$ and $\lambda_H = 1 - \lambda_L$.

can lead to an inferior outcome for everyone, even if the use of such experts is optional.

Closest to our setting is the work of Lingenbrink and Iyer (2019), who study optimal signaling for services with unobservable queues. Even though our work builds on the machinery developed in Lingenbrink and Iyer (2019), there are also key differences, which we discuss next. Lingenbrink and Iyer (2019) are concerned with maximizing the service provider's revenue using information sharing as well as static pricing. As such, the goal of an optimal signaling mechanism in that setting is to persuade more customers to join the queue. However, in our setting, the service provider uses an information sharing mechanism to improve welfare outcomes by persuading more low-need users to leave. Further, Lingenbrink and Iyer (2019) mainly focus on a setting with homogeneous users, whereas we study a setting with different user types. Relatedly, Anunrojwong et al. (2019) study persuasion of non-expected-utility maximizing agents and apply it to study throughput maximization in queues where customers' disutility depends on the variance of their waiting times.

Finally, Das et al. (2017) also study how optimal information sharing mechanisms can reduce congestion in a traffic network when a user chooses a path among a set of paths, some of which have uncertain states. In particular, the authors consider a static setting where a continuum of users simultaneously decide on the path they wish to take to minimize their own cost and show that all public signaling mechanisms yield the same outcomes as full information (or no information). Our paper complements this work by considering a dynamic setting in which users of different types sequentially arrive over time. Upon the arrival of each user, the service provider sends a state-dependent signal. We show that public signaling can be effective in improving welfare outcomes when compared with special mechanisms of full information and no information.

1.4.2. Strategic Behavior in Queueing Systems. Following the seminal work of Naor (1969), a stream of literature has focused on analyzing queuing systems where users are strategic. (See the surveys by Hassin 2016 and Ibrahim 2018, and the references therein.) In particular, Hassin and Koshman (2017) study mechanisms for profit maximization in an $M/M/1$ queue with homogeneous customers and establish the optimality of an information sharing mechanism that, along with appropriate prices, makes "high-low" announcements where arriving customers receive a "low" announcement if and only if the queue length is below Naor's socially optimal threshold. Our work differs in two main respects. First, as our application context is social services, our model ignores pricing as a lever (effectively taking prices as exogenous) but optimizes over all information-sharing mechanisms. Second, our objective is Pareto improvement of (customer) welfare rather than profit maximization, the two objectives usually being opposed. Focusing on the information-sharing aspect (for an unobservable queue), Allon et al. (2011) consider a cheap talk setting where the service provider does not have commitment power. Additionally, as discussed above, Lingenbrink and Iyer (2019) consider information design in conjunction with pricing in order to maximize the service provider's revenue. Finally, Che and Tercieux (2020) study the optimal design of a queuing system where the planner decides on several aspects, including the queue discipline, entry, abandonment, and information sharing. Interestingly, they show that the optimal design is to follow first-come-first-serve, recommend that users join up to a threshold (in queue size), and

never recommend abandonment for a user in the queue.

1.4.3. Dynamic Allocation of Social Goods. Our paper is also related to the literature on the dynamic allocation of social goods such as public housing (Kaplan 1984) and donated organs (Ashlagi et al. 2013, 2019). Recently, Leshno (2017) and Arnosti and Shi (2020) consider settings where the user has a heterogeneous preference over arriving goods and, thus, she faces a trade-off between waiting longer and accepting a less preferred good.[9] (A similar trade-off exists in dynamic matching, as studied in Doval and Szentes 2018 and Baccara et al. 2020.) These papers focus on designing efficient allocation mechanisms such as waitlist mechanisms. We complement this literature by studying the role that information sharing can play in improving welfare for social services. Finally, the recent work of Ashlagi et al. (2020) studies the dynamic allocation of heterogeneous items, where an agent's value for an item is pair-specific, that is, jointly depends on the agent type and the quality of the item. The information design aspect of Ashlagi et al. (2020) differs from ours in that it is concerned with information disclosure about the (unobservable) quality of an arriving item. In contrast, in this paper, we assume that the service rate is known to the user, and we focus on the information disclosure with regard to the (unobservable) congestion level.

1.4.4. Mechanism Design Without Money. Our work investigates the power of information design to reduce congestion in social services, where the usage of monetary payments to shape agents' incentives is either impractical or unpalatable. As such, it is broadly related to the growing literature on mechanism design without money. Motivated by wide-ranging applications, this stream of literature studies resource allocation without relying on monetary payments. For examples of static settings, see Procaccia and Tennenholtz (2009), Prendergast (2017); dynamic settings are studied in Balseiro et al. (2019), Gorokh et al. (2021).

2. Model

In the following, we describe a model of information design for improving welfare outcomes in a queueing setting with heterogeneous users. Our model builds upon that of Lingenbrink and Iyer (2019), who study revenue maximization in a related queueing setting.

Consider a service provider who provides a social service to a stream of users arriving over time. Due to capacity constraints, the arriving users possibly wait in an unobservable queue for service, where they are served on a first-come-first-serve basis by a single server. Each user's service time is independent and identically distributed as an exponential distribution with rate 1.[10]

Arriving users must decide whether to join the queue and wait for the service or to leave for an outside option. Upon joining, we assume that there is no abandonment: if a user joins the queue, then she will stay until service completion. To describe users' utility, we start with discussing their outside options. We model the users as belonging to one of two groups that differ in the quality of their outside options. Specifically, we assume that each user is either a (1) *high-need* user, who has no viable outside option, which we model by letting their utility for taking the outside option be $-\infty$; or a (2) *low-need* user who has a viable outside option whose utility we normalize to 0. We denote a user's type as H if they are high-need, and by L if they are low-need. We assume that users of type $i \in \{\mathsf{H},\mathsf{L}\}$ arrive according to an independent Poisson process with rate λ_i, with $\lambda = \lambda_{\mathsf{L}} + \lambda_{\mathsf{H}}$ denoting the total arrival rate. To avoid trivialities, we assume that $\lambda_{\mathsf{L}} > 0$. In our analysis, we also assume that $\lambda \le 1$, to capture the setting where the social service is not undercapacitated.

On joining the queue to obtain service, each user receives a net utility composed of the benefit from the social service and a cost of waiting until service completion. Formally, the utility function of a type $i \in \{\mathsf{L},\mathsf{H}\}$ user is given by $u_i : \mathbb{N}_0 \to \mathbb{R}$, where $u_i(n)$ denotes her utility on joining a queue with n users already in the system, either in queue or being served.[11] We make the natural assumptions that $u_i(0) > 0$, and $\lim_{k\to\infty} u_i(k) < 0$. Further we make the following assumption.

Assumption 1 (Positive and Diminishing Incremental Waiting Costs). *The utility functions satisfy the following monotonicity assumptions:*

1. *For each type $i \in \{\mathsf{H},\mathsf{L}\}$, the utility function $u_i(n)$ is strictly decreasing in n.*
2. *The difference $u_{\mathsf{L}}(n) - u_{\mathsf{L}}(n+1)$ is nonincreasing in n.*

We remark that the monotonicity assumption on the utility of both types is natural, and it reflects the fact that waiting for service completion imposes a waiting cost on the users. The second condition requires that while each additional user ahead in queue imposes greater waiting costs on a L-type user, the incremental cost decreases with more users ahead in queue. We note that the linear utility function, that is, $u_{\mathsf{L}}(n) = 1 - c(n+1)$ for some $c > 0$, satisfies both conditions.

We assume that the users are strategic and Bayesian in their joining decisions. Because high-need users have no viable outside option, any such arriving user always joins the queue for service. On the other hand, the low-need users may decide to leave for the outside option, based on their beliefs about the queue state. Since the queue is unobservable to the users, the service provider seeks to leverage his informational advantage to influence the low-need users' decision, with the goal toward

improving welfare outcomes. To that end, the service provider commits to a *signaling mechanism* as follows: the service provider selects a set of possible signals $\mathbb{S}$, and a mapping $\sigma : \mathbb{N}_0 \times \mathbb{S} \to [0,1]$, such that, if there are n users already in queue upon the arrival of a user, then he sends a signal $s \in \mathbb{S}$ to the user with probability $\sigma(n,s) \in [0,1]$. (We require $\sum_{s\in\mathbb{S}}\sigma(n,s) = 1$ for all n.) Note that since high-need users in our model have no viable outside option and, hence, always join the queue, the service provider can implement a signaling mechanism without the knowledge of user types.

Given the signaling mechanism, we require the low-need users' choices to constitute an equilibrium. Informally, the equilibrium requires that, in the steady state that arises from the users' actions, each low-need user is acting optimally. To elaborate further, given the steady-state distribution π, we require that a low-need user joins the queue upon receiving a signal $s \in \mathbb{S}$ if and only if her expected utility from joining $\mathbf{E}_\pi[u_\mathsf{L}(n)|s]$ is greater than 0, the utility of her outside option. (We assume that ties are broken in favor of joining; we note that, due to the negative externalities that users in the queue impose on each other, the welfare under other tie-breaking rules can only be better.) Note that the steady-state distribution π itself is determined endogenously in equilibrium from the users' actions. To avoid unnecessary notational burden, we refrain from formally defining the equilibrium for general signaling mechanisms and point the reader to Lingenbrink and Iyer (2019). Instead, using standard arguments based on the revelation principle (see, e.g., Bergemann and Morris 2016, Candogan and Drakopoulos 2019, Lingenbrink and Iyer 2019), one can show that it suffices to consider *obedient* binary signaling mechanisms. These are the mechanisms where the signals are limited to "join" and "leave"—which we represent as 1 and 0, respectively—and for which in the resulting user equilibrium a high-need user always joins, and a low-need user joins upon receiving signal 1 and leaves otherwise. We describe such mechanisms more formally next.

First note that a binary signaling mechanism can be described by $\{p_n : n \geq 0\}$, where p_n denotes the probability that an L-type user receives the signal $s = 1$ ("join"), when the queue length is n upon her arrival. Assuming that all users follow their recommendation, let $\pi = \{\pi_n : n \geq 0\}$ denote the resulting steady-state distribution. By elementary queueing theory, the steady-state distribution satisfies the following detailed-balance conditions (Gross et al. 2018):

$$\pi_{n+1} = (\lambda_\mathsf{L} p_n + \lambda_\mathsf{H})\pi_n, \quad \text{for all } n \geq 0. \tag{1}$$

Given the steady-state distribution and using Bayes's rule, an arriving L-type user receiving the signal $s = 1$ ("join") believes the queue length is $n \geq 0$ with probability $\pi_n p_n / \sum_{k\in\mathbb{N}_0}\pi_k p_k$. Similarly, an arriving L-type user receiving the signal $s = 0$ ("leave") believes the queue length is $n \geq 0$ with probability $\pi_n(1-p_n)/\sum_k \pi_{k\in\mathbb{N}_0}(1-p_k)$.

For an L-type user, let $U_\mathsf{L}(s,a)$ denote her expected utility upon receiving a signal $s \in \{0,1\}$ and choosing an action $a \in \{\mathsf{join}, \mathsf{leave}\}$. Note that we have $U_\mathsf{L}(s, \mathsf{leave}) = 0$. (Recall that L-type's outside option is normalized to 0.) On the other hand, we have

$$U_\mathsf{L}(1,\mathsf{join}) = \sum_{n\in\mathbb{N}_0} \frac{\pi_n p_n}{\sum_{k\in\mathbb{N}_0}\pi_k p_k} u_\mathsf{L}(n) = \frac{\sum_{n\in\mathbb{N}_0}(\pi_{n+1} - \lambda_\mathsf{H}\pi_n)u_\mathsf{L}(n)}{\sum_{n\in\mathbb{N}_0}(\pi_{n+1} - \lambda_\mathsf{H}\pi_n)},$$

$$U_\mathsf{L}(0,\mathsf{join}) = \sum_{n\in\mathbb{N}_0} \frac{\pi_n(1-p_n)}{\sum_{n\in\mathbb{N}_0}\pi_k(1-p_k)} u_\mathsf{L}(n) = \frac{\sum_{n\in\mathbb{N}_0}(\lambda\pi_n - \pi_{n+1})u_\mathsf{L}(n)}{\sum_{n\in\mathbb{N}_0}(\lambda\pi_n - \pi_{n+1})}$$

Here, the second equality in each line follows from the fact that $\lambda_\mathsf{L}\pi_n p_n = \pi_{n+1} - \lambda_\mathsf{H}\pi_n$ and $\lambda_\mathsf{L}\pi_n(1-p_n) = \lambda\pi_n - \pi_{n+1}$, which follow from the detailed-balance condition in Equation (1).

In an obedient binary signaling mechanism, an L-type user must find it incentive-compatible to follow the service provider's recommendations. Thus, in such a mechanism, we must have the following *obedience* constraints: $U_\mathsf{L}(1,\mathsf{join}) \geq U_\mathsf{L}(1,\mathsf{leave}) = 0$ and $U_\mathsf{L}(0,\mathsf{join}) \leq U_\mathsf{L}(0,\mathsf{leave}) = 0$. This, in turn, yields the following constraints on the steady-state distribution π:

$$J(\pi) \triangleq \sum_{n=0}^{\infty}(\pi_{n+1} - \lambda_\mathsf{H}\pi_n)u_\mathsf{L}(n) \geq 0, \tag{JOIN}$$

$$L(\pi) \triangleq \sum_{n=0}^{\infty}(\lambda\pi_n - \pi_{n+1})u_\mathsf{L}(n) \leq 0 \tag{LEAVE}$$

Using the preceding constraints, the following result, from Lingenbrink and Iyer (2019), establishes a correspondence between obedient binary signaling mechanisms and a set of all distributions satisfying obedience constraints. We omit the proof for brevity.

Lemma 1 (Lingenbrink and Iyer 2019). *For any obedient binary signaling mechanism, the steady-state distribution π satisfies the following conditions:*

1. *Distributional constraints: $\sum_{n\in\mathbb{N}_0}\pi_n = 1$ and $\pi_n \geq 0$ for all $n \geq 0$;*
2. *Detailed-balance constraints: $\lambda_\mathsf{H}\pi_n \leq \pi_{n+1} \leq (\lambda_\mathsf{H} + \lambda_\mathsf{L})\pi_n$ for all $n \in \mathbb{N}_0$; and*
3. *Obedience constraints (JOIN) and (LEAVE) as defined above.*

Conversely, for any distribution π satisfying the preceding sets of constraints, there exists an obedient binary signaling mechanism $\{p_n : n \geq 0\}$, with $p_n = \frac{\pi_{n+1} - \lambda_\mathsf{H}\pi_n}{\lambda_\mathsf{L}\pi_n}$ whenever $\pi_n > 0$ (and arbitrary otherwise).

We let Π_{SM} denote the set of all distributions that satisfy the three sets of constraints mentioned above. (Here, SM stands for *signaling mechanism*.) Here, the second constraints arise from the detailed-balance

conditions in Equation (1) and the fact that $p_n \in [0,1]$ for all n.

In addition to simplifying notation, the preceding result enables us to describe the user welfare in an obedient binary signaling mechanism purely in terms of the resulting distribution $\pi \in \Pi_{\mathsf{SM}}$. In particular, for any $\pi \in \Pi_{\mathsf{SM}}$, the welfare of type i users, denoted by $W_i(\pi)$, is given by

$$W_{\mathsf{L}}(\pi) = \lambda_{\mathsf{L}} \sum_{n=0}^{\infty} \pi_n \left(\frac{\pi_{n+1} - \lambda_{\mathsf{H}} \pi_n}{\lambda_{\mathsf{L}} \pi_n} \right) u_{\mathsf{L}}(n)$$

$$= \sum_{n=0}^{\infty} (\pi_{n+1} - \lambda_{\mathsf{H}} \pi_n) u_{\mathsf{L}}(n) = J(\pi), \tag{2}$$

$$W_{\mathsf{H}}(\pi) = \lambda_{\mathsf{H}} \sum_{n=0}^{\infty} \pi_n u_{\mathsf{H}}(n). \tag{3}$$

Here, the first line follows from the fact that the arrival rate of L-type users is λ_{L} and that if the queue length is n, which occurs with probability π_n in steady state, then an arriving L-type user joins the queue with probability $(\pi_{n+1} - \lambda_{\mathsf{H}}\pi_n)/\lambda_{\mathsf{L}}\pi_n$ and receives utility $u_{\mathsf{L}}(n)$. Similarly, the second line follows from the fact that an H-type user always joins upon arrival.

Since we focus on a social service setting, we seek to understand the effectiveness of information design in improving the welfare outcomes for *both* types. In this context, we use the following definition of *Pareto efficiency*.

Definition 1 (Pareto Efficiency). For any two $\pi, \hat{\pi} \in \Pi_{\mathsf{SM}}$, we say that $\hat{\pi}$ *Pareto-dominates* π, if $W_i(\hat{\pi}) \geq W_i(\pi)$ for $i \in \{\mathsf{L}, \mathsf{H}\}$ with a strict inequality for at least one i. Further, we say that a distribution $\pi \in \Pi_{\mathsf{SM}}$ is *Pareto-efficient* within the class Π_{SM} if and only if there exists no $\hat{\pi} \in \Pi_{\mathsf{SM}}$ that Pareto-dominates π.

Hereafter, we frequently abuse the terminology to say that an obedient binary signaling mechanism is Pareto-efficient (within the class of such mechanisms), if the corresponding steady-state distribution (as per Lemma 1) is Pareto-efficient[12] within the class Π_{SM}.

For our comparative analysis, we look at two specific signaling mechanisms that capture the two extremes of information sharing:

(a) Full-information mechanism (fi): Here, the service provider always reveals the queue length to an arriving user. Consequently, L-type users join the queue at all queue lengths k with $u_{\mathsf{L}}(k) \geq 0$. Letting m_{fi} denote the smallest integer k for which $u_{\mathsf{L}}(k) < 0$, the corresponding steady-state distribution π^{fi} satisfies $\pi^{\mathsf{fi}}_{n+1} = \lambda \pi^{\mathsf{fi}}_n$ for $n < m_{\mathsf{fi}}$, and $\pi^{\mathsf{fi}}_{n+1} = \lambda_{\mathsf{H}} \pi^{\mathsf{fi}}_n$ otherwise.

(b) No-information mechanism (ni): Here, the service provider reveals no information to the users. Consequently, the strategy of an arriving L-type user is independent of n. Letting p^{ni} denote the probability with which a user joins the queue in a symmetric equilibrium, the corresponding steady-state distribution π^{ni} satisfies $\pi^{\mathsf{ni}}_n = (p^{\mathsf{ni}}\lambda_{\mathsf{L}} + \lambda_{\mathsf{H}})^n \pi^{\mathsf{ni}}_0$ for all $n \geq 0$. In Lemma EC.2 (in Section EC.2 or the e-companion), we characterize the joining probability p^{ni} in an equilibrium.

In the following, we also consider *admission policies*, where the service provider can enforce the joining or leaving of any user, regardless of the user's type or her incentives. Whereas such enforcement is clearly practically infeasible, it serves as a benchmark against which the welfare outcomes of signaling mechanisms can be compared. Formally, an admission policy can be described by the class of distributions Π_{AP} that satisfy the distributional and the detailed-balance constraints from Lemma 1, but it need not satisfy the obedience constraints (JOIN) *and* (LEAVE). (Here, AP stands for *admission policy*.) Analogous to Definition 1, we define Pareto-efficiency within the class Π_{AP}. Observe that $\Pi_{\mathsf{SM}} \subseteq \Pi_{\mathsf{AP}}$; that is, any signaling mechanism is also an admission policy (one that also respects user incentives), and, hence, any signaling mechanism $\pi \in \Pi_{\mathsf{SM}}$ that is Pareto-efficient within Π_{AP} is also Pareto-efficient within Π_{SM}, but the converse may not hold. This observation motivates our choice of Π_{AP} as a welfare benchmark.

Before we end this section, we note that both Π_{AP} and Π_{SM} are closed and convex, and the welfare functions as defined in (2) and (3) are linear in π. Thus, the sets $\{(W_{\mathsf{L}}(\pi), W_{\mathsf{H}}(\pi)), \pi \in \Pi_{\mathsf{SM}}\}$ and $\{(W_{\mathsf{L}}(\pi), W_{\mathsf{H}}(\pi)), \pi \in \Pi_{\mathsf{AP}}\}$ are also convex. As a consequence, it follows that any $\hat{\pi}$ that is Pareto-efficient within the class of signaling mechanisms Π_{SM} (or admission policies Π_{AP}), is a solution to the (linear) optimization problem that maximizes the convex combination of the two user types' welfare over Π_{SM} (respectively, Π_{AP}). In particular, let $W(\pi, \theta) \triangleq \theta W_{\mathsf{L}}(\pi) + (1-\theta) W_{\mathsf{H}}(\pi)$ for all $\pi \in \Pi_{\mathsf{AP}}$ and for $\theta \in [0,1]$. Then, each Pareto-efficient signaling mechanism maximizes $W(\pi, \theta)$ over Π_{SM} for some $\theta \in [0,1]$, and each maximizer of $W(\pi, \theta)$ over Π_{SM} for $\theta \in (0,1)$ is a Pareto-efficient signaling mechanism (Mas-Colell et al. 1995, proposition 16.E.2). Similarly, any Pareto-efficient admission policy is a solution to $\max_{\pi \in \Pi_{\mathsf{AP}}} W(\pi, \theta)$ for some $\theta \in [0,1]$ (and each maximizer of $W(\pi, \theta)$ over Π_{AP} for $\theta \in (0,1)$ is a Pareto-efficient admission policy).[13] Furthermore, for any Pareto-efficient π, the specific $\theta \in [0,1]$ for which π maximizes $W(\cdot, \theta)$ captures the relative importance that the service provider ascribes to improving the welfare of the two types. In this context, for a given θ, we refer to the admission policy that achieves the maximum as the *first-best*.

3. Structural Characterization

In this section, we provide structural characterizations of the Pareto-efficient signaling mechanisms and admission policies. We use these structural characterizations in Sections 4 and 5 to evaluate the effectiveness

of signaling mechanisms in improving welfare outcomes, and we compare their performance against admission policies and simple signaling mechanisms.

Before we begin, for the sake of completeness, we state the following technical result that establishes the existence of Pareto-efficient signaling mechanisms and admission policies. The proof follows from the observation that the sets Π_{AP} and Π_{SM} (or closures of some relevant subsets) are weakly compact, and, hence, the maximizers of $W(\pi,\theta)$ over these sets exist for all $\theta\in[0,1]$. It is straightforward to show that these maximizers are Pareto-efficient within their respective class. The formal proof is provided in Section EC.1 of the e-companion.

Lemma 2 (Existence). *For $\lambda<1$ and for each $\theta\in[0,1]$, there exists a signaling mechanism (admission policy) that maximizes $W(\pi,\theta)$ over all $\pi\in\Pi_{\mathsf{SM}}$ (respectively, $\pi\in\Pi_{\mathsf{AP}}$). If further $\lim_{n\to\infty}u_i(k)=-\infty$ for each $i\in\{\mathsf{L},\mathsf{H}\}$, then the result also holds for $\lambda=1$.*

Next, we define the following threshold structure among distributions $\pi\in\Pi_{\mathsf{AP}}$.

Definition 2 (Threshold Structure). We say that a given $\pi\in\Pi_{\mathsf{AP}}$ has a *threshold* structure if there exists an $m\in\mathbb{N}_0\cup\{\infty\}$, such that $\pi_{k+1}=\lambda\pi_k$ for all $k<m$, and $\pi_{k+1}=\lambda_{\mathsf{H}}\pi_k$ for all $k>m$. In such a setting, we say that the distribution π has a threshold $x=m+a\in\mathbb{R}_+$, where $a=(\pi_{m+1}-\lambda_{\mathsf{H}}\pi_m)/\lambda_{\mathsf{L}}\pi_m\in[0,1]$.

Informally, a distribution $\pi\in\Pi_{\mathsf{AP}}$ has a threshold structure with threshold equal to $x=m+a\in[m,m+1]$, if an arriving L-type user is asked to join the queue with probability 1 for all queue lengths strictly less than m, asked to leave with probability 1 for all queue lengths strictly greater than m, and asked to join the queue with probability $a\in[0,1]$ if the queue length is exactly m. (Note that a threshold ∞ corresponds to the case where an arriving L-type user is asked to join regardless of the queue length.)

Our first result states that any Pareto-efficient signaling mechanism has a threshold structure. The proof, which is deferred to Section EC.1 of the e-companion, follows from a perturbation analysis similar to that in Lingenbrink and Iyer (2019): we show that given any $\pi\in\Pi_{\mathsf{SM}}$ that does not have a threshold structure, one can perturb it to obtain a $\hat{\pi}\in\Pi_{\mathsf{SM}}$ that Pareto-dominates it.

Theorem 1 (Threshold Structure of Pareto-Efficient Signaling Mechanisms). *Any signaling mechanism $\pi\in\Pi_{\mathsf{SM}}$ that is Pareto-efficient within the class Π_{SM} has a threshold structure, with the threshold less than or equal to the full-information threshold m_{fi}.*

Furthermore, using the same argument as above, we obtain that the Pareto-efficient admission policies also have a threshold structure. We omit the proof for brevity.

Theorem 2 (Threshold Structure of Pareto-Efficient Admission Policies). *Any admission policy $\pi\in\Pi_{\mathsf{AP}}$ that is Pareto-efficient within the class Π_{AP} has a threshold structure, with the threshold less than or equal to the full-information threshold m_{fi}.*

Having established the threshold structure of any Pareto-efficient distribution (either within Π_{AP} or Π_{SM}), we next state another key structural property of Pareto-efficient distributions within Π_{SM}. The result implies that in any Pareto-efficient $\pi\in\Pi_{\mathsf{SM}}$ that is not Pareto-efficient within Π_{AP}, the obedience constraint that binds is the constraint (LEAVE). Put differently, it is the (LEAVE) condition that acts as a hurdle for a Pareto-efficient admission policy to be implementable as a signaling mechanism. The intuition behind this result lies in the observation, common in many congested service systems, that L-type users do not internalize the negative externalities they impose on other users (both L-type and H-type) by joining the queue. Hence, the L-type users are naturally more inclined to join the queue than leave, and the challenge in information sharing is in ensuring that when the L-type users are asked to leave, they find it incentive-compatible to do so. The proof of this theorem is also deferred to Section EC.1 of the e-companion.

Theorem 3 (Significance of the (LEAVE) Hurdle). *Suppose that, for a signaling mechanism $\pi\in\Pi_{\mathsf{SM}}$, the obedience constraint* (LEAVE) *does not bind; that is, $L(\pi)<0$. Then, π is Pareto-efficient within the class Π_{SM} of signaling mechanisms if and only if it is Pareto-efficient within the class Π_{AP} of admission policies.*

In concluding this section, we note that the preceding result raises the intriguing possibility of the existence of a signaling mechanism that is Pareto-efficient not only within the class Π_{SM} of signaling mechanisms, but also within the broader class Π_{AP} of admission policies. For any such mechanism, it follows that, under a practically infeasible setting where the service provider observes the types of users and is allowed to enforce the joining and leaving of users, he cannot jointly improve both types' welfare. Put differently, the existence of such mechanisms also implies the existence of admission policies where the L-type users' incentive constraints are satisfied for "free."

A trivial instance of such a scenario can arise, for example, in cases where λ_{H} is large enough, and the admission policy always bars L-type users from joining the queue. First, such an admission policy must be Pareto-efficient, as any other policy that lets some L-type users in would necessarily reduce the welfare of the H-type users. Furthermore, such an admission policy can be implemented as a no-information mechanism, which satisfies obedience constraints, as congestion in the queue with just the H-type users makes joining undesirable for

the L-type users. Excluding such trivial scenarios, a natural question is whether there exist signaling mechanisms that do not exclude any types but are still Pareto-efficient within the class of admission policies Π_{AP}. In Section 6, we present numerical examples where such mechanisms indeed exist (see Figure 4 and its related discussion).

4. Mechanism Comparisons

Having characterized the structure of Pareto-efficient signaling mechanisms, we compare such mechanisms against various benchmarks in two different settings. First, in Section 4.1, we consider the homogeneous setting where all users have L-type, that is, $\lambda_{\mathsf{H}} = 0$.[14] Then, in Section 4.2, we consider the heterogeneous setting with both types of users present. As we discuss below, the two settings exhibit striking contrast in the effectiveness of signaling mechanisms for welfare improvement.

In light of Theorems 1 and 2, for ease of presentation, we use the following simplified notations for a threshold policy: For $x \in \mathbb{R}_+$, the threshold policy x is an admission policy that gives rise to a steady-state distribution $\pi_x \in \Pi_{\mathsf{AP}}$ that has a threshold structure, as defined in Definition 2, with threshold x. For such a policy, with a slight abuse of notation, for $i \in \{\mathsf{L}, \mathsf{H}\}$ we denote $W_i(\pi_x)$ simply by $W_i(x)$. We do the same for $J(\pi_x)$ and $L(\pi_x)$.

4.1. Homogeneous Users

We start our comparative studies by analyzing the special case of homogeneous users ($\lambda_{\mathsf{H}} = 0$). Observe that in this single-type setting, Pareto dominance is equivalent to optimality, in terms of maximizing the welfare of L-type users. Consequently, we let sm denote the *optimal* signaling mechanism, the one that maximizes the welfare of L-type users.

In the following proposition, we compare the optimal signaling mechanism sm with full information (fi). With a slight abuse of notation, for $\mu \in \{\mathsf{fi}, \mathsf{sm}\}$, we denote by $W_{\mathsf{L}}(\mu)$ the L-type welfare under mechanism μ. We have the following result, whose proof is provided in Section EC.3 of the e-companion.

Proposition 1 (Limits of Information Design). *In the homogeneous setting, we have*

$$W_{\mathsf{L}}(\mathsf{fi}) \geq \beta_{\mathsf{fi}} \cdot W_{\mathsf{L}}(\mathsf{sm}),$$

where $\beta_{\mathsf{fi}} \triangleq (\sum_{n=0}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{L}}^n)/(\sum_{n=0}^{m_{\mathsf{fi}}} \lambda_{\mathsf{L}}^n) \geq 1 - \frac{1}{m_{\mathsf{fi}}+1}$, *and* m_{fi} *is the full-information threshold. Further, the equality holds; that is,* $W_{\mathsf{L}}(\mathsf{fi}) = W_{\mathsf{L}}(\mathsf{sm})$ *if and only if* $W_{\mathsf{L}}(m_{\mathsf{fi}}) \geq W_{\mathsf{L}}(m_{\mathsf{fi}} - 1)$.

The preceding proposition states that in the homogeneous setting, signaling mechanisms are not very effective in improving the welfare beyond that already achieved by the full-information mechanism. To gain some intuition, observe that, in general Bayesian persuasion settings, the performance gains are typically achieved by pooling, in the persuaded agents' beliefs, the "good" and the "bad" states of the system. However, in a queueing setting, the linear nature of the underlying Markovian system precludes any such simple pooling of states in the agents' belief: the only way for the system to reach a bad state (one with a long queue length) is by progressing through all intermediate queue lengths. Because of this, agents are not easily persuaded. Formally, the proof proceeds by showing that a threshold mechanism with threshold $x < m_{\mathsf{fi}} - 1$ will not be incentive-compatible. In particular, we will show that if $x < m_{\mathsf{fi}} - 1$, then the second obedience constraint (LEAVE) will be violated. This can also be intuitively explained: if the threshold were below $m_{\mathsf{fi}} - 1$, then the queue will never be longer than m_{fi}, and any user receiving a "leave" signal will realize this and will want to join the queue, thus violating the (LEAVE) condition. Since we have already shown (Theorem 1) that the threshold of the signaling mechanism is at most m_{fi}, we conclude that the threshold of the signaling mechanism is between $m_{\mathsf{fi}} - 1$ and m_{fi}. Thus, any small improvement in welfare of sm over fi stems from the difference in user behavior when the queue length is $m_{\mathsf{fi}} - 1$, where users always join under fi but may sometimes leave under sm. Building on this observation, in the proof we bound the relative welfare gain by a factor $1/\beta_{\mathsf{fi}}$.

In contrast, in a "sufficiently" heterogeneous population, the presence of H-type users makes persuading the L-type ones possible: the queue now consists of two types of users. Therefore, a user's belief about the queue length will now depend on the behavior of both types. Leveraging this, a signaling mechanism can set a threshold much lower than that of the full-information mechanism without violating the (LEAVE) condition. This, in turn, can result in substantial welfare gain. (For numerical examples, see Figure 2 and its related discussion in Section 6.)

Finally, in the following proposition (proved in Section EC.3 of the e-companion), we show that even though information design results in limited or no improvement over the full-information mechanism, it always outperforms the no-information mechanism.

Proposition 2 (Suboptimality of the No-Information Mechanism). *In the homogeneous setting, the no-information mechanism is never optimal. In particular, the welfare under the no-information mechanism* ni *satisfies* $W_{\mathsf{L}}(\mathsf{ni}) < (1 - \lambda_{\mathsf{L}}^{m_{\mathsf{fi}}+1}) W_{\mathsf{L}}(\mathsf{fi}) \leq (1 - \lambda_{\mathsf{L}}^{m_{\mathsf{fi}}+1}) W_{\mathsf{L}}(\mathsf{sm})$.

4.2. Heterogeneous Users

Next, we proceed to a setting where the population is a mixture of the two types. Here, we have two objectives, namely, the welfare of both types. As discussed before, to examine the effectiveness of information design in

improving the welfare of both types, we focus on the notion of Pareto efficiency. In the following, we draw comparisons between Pareto-efficient optimal signaling mechanisms with the extreme forms of sharing information, namely, full information and no information.

Our main result in this section is the following proposition, which provides the necessary and sufficient conditions on arrival rates under which the full-information and no-information mechanisms are Pareto-dominated.

Proposition 3 (Power of Information Design). *The following statements hold.*

1. *For any $\lambda_{\mathsf{H}} > 0$, there exists a signaling mechanism that Pareto-dominates the full-information mechanism if and only if $\lambda_{\mathsf{L}} \in [\bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}}), 1)$, where $\bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}}) \in (0, 1-\lambda_{\mathsf{H}}]$ and we have*

$$\bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}}) \geq (1-\lambda_{\mathsf{H}}) \frac{u_{\mathsf{L}}(m_{\mathsf{fi}}-1)}{u_{\mathsf{L}}(0) - \sum_{k=1}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{H}}^k (u_{\mathsf{L}}(k-1) - u_{\mathsf{L}}(k))} > 0.$$

2. *For $\lambda < 1$ and $\lambda_{\mathsf{L}} \in [\bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}}), 1)$, if the utility functions are such that $u_{\mathsf{L}}(m_{\mathsf{fi}}-1) \leq W_{\mathsf{L}}(\mathsf{fi})$, $u_{\mathsf{H}}(m_{\mathsf{fi}}) \leq W_{\mathsf{H}}(\mathsf{fi})$, $L(m_{\mathsf{fi}}-1) \leq 0$, and $W_{\mathsf{H}}(\mathsf{fi}) > 0$, then we have*

$$W_{\mathsf{L}}(\mathsf{sm}) \geq \beta_{\mathsf{L,sm}} \cdot W_{\mathsf{L}}(\mathsf{fi}) \quad \text{and} \quad W_{\mathsf{H}}(\mathsf{sm}) \geq \beta_{\mathsf{H,sm}} \cdot W_{\mathsf{H}}(\mathsf{fi}),$$

where $\beta_{\mathsf{L,sm}} \triangleq \left(1 + \frac{\lambda_{\mathsf{H}}(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m_{\mathsf{fi}}-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m_{\mathsf{fi}}-1}}\right) > 1$ *and* $\beta_{\mathsf{H,sm}} \triangleq \left(1 + \frac{(1-\lambda_{\mathsf{H}})(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m_{\mathsf{fi}}-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m_{\mathsf{fi}}-1}}\right) > 1.$

3. *The no-information mechanism* ni *is Pareto-dominated by a signaling mechanism if and only if $\lambda_{\mathsf{H}} \in [0, \bar{\Lambda}_{\mathsf{H}})$, where $\bar{\Lambda}_{\mathsf{H}}$ is the unique root in (0, 1) of the function $g(x) \triangleq \sum_{k\in\mathbb{N}_0}(1-x)x^k u_{\mathsf{L}}(k)$. Here, $\bar{\Lambda}_{\mathsf{H}}$ is the smallest arrival rate of the* H*-type users under which no* L*-type user joins the queue in the equilibrium of the no-information mechanism.*

The preceding result states that, as long as the arrival rate λ_{L} of L-type users is sufficiently high, the welfare of both types can be improved using information design, as compared with full-information sharing. As we discuss below (after stating Theorem 4), under sufficiently high λ_{L}, the negative externality that a L-type user imposes on other L-type users exceeds the utility she receives from the service; in such settings, not revealing all the information about the state helps the L-type users to internalize this negative externality. To illustrate the benefit of information design in such a regime, the second part of the proposition places a lower bound on the welfare gain of each type under certain conditions on the utility functions of the two types. On the other hand, as long as the arrival rate λ_{H} of H-type users is not too high, information design can improve the welfare over no-information sharing. In this case, information design helps by providing enough state information to correlate L-type users' actions with the queue state. Taken together, the result implies that information design has an unambiguously positive role for welfare improvement in settings where the type composition of user population is fairly balanced.

We defer the complete proof of Proposition 3 to Section EC.3 of the e-companion. The proof relies on two intermediate results that have a similar dichotomous structure, characterizing when each of the two benchmarks is Pareto-efficient. We devote the rest of this section to discussing (and proving) the two results, and their relation to our main proposition. The proofs of both results are given in Section EC.3 of the e-companion.

The first result presents the following dichotomy for the full-information benchmark.

Theorem 4 (Information Design vs. Full Information). *Exactly one of the following two statements holds:*

1. *The full-information mechanism* fi *is Pareto-efficient within the class of admission policies.*

2. *There exists a signaling mechanism that Pareto-dominates the full-information mechanism.*

Further, the first case occurs if and only if $W_{\mathsf{L}}(m_{\mathsf{fi}}) > W_{\mathsf{L}}(m_{\mathsf{fi}}-1)$.

To understand the implications of the preceding result, consider an admission policy with threshold $x < m_{\mathsf{fi}}$. As x increases, more L-type users are served by the service provider, increasing their utility. At the same time, the negative externality that each L-type user imposes on other L-type users increases as x increases. (This is in addition to the negative externalities imposed on H-type users.) The preceding result states that for the full-information mechanism to be Pareto-efficient, the gains from serving more L-type users must dominate the negative externalities that they impose on other L-type users (which is succinctly captured by the condition $W_{\mathsf{L}}(m_{\mathsf{fi}}) > W_{\mathsf{L}}(m_{\mathsf{fi}}-1)$). Conversely, if serving more L-type users imposes greater negative externality on other L-type users, then our result states that information design can leverage this to improve the welfare of *both* types over the full-information mechanism. Finally, tying back to Proposition 3, for the effect of negative externality to dominate, the arrival rate λ_{L} of L-type users must be large enough, as captured by the condition $\lambda_{\mathsf{L}} \geq \bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}})$.

Next, we obtain the following dichotomy for the no-information benchmark.

Theorem 5 (Information Design vs. No Information). *Exactly one of the following two statements holds:*

1. *The no-information mechanism* ni *is Pareto-efficient within the class of admission policies.*

2. *There exists a signaling mechanism that Pareto-dominates the no-information mechanism.*

Further, the first case occurs if and only if all L*-type users choose their outside option under the no-information mechanism.*

The preceding result neatly breaks the analysis into two cases. In the first case, even if no other L-type users join the queue, the outside option is more desirable for a L-type user. In other words, an L-type user does not need much persuasion to forgo the social service and avail the outside option. In such instances, any information shared by the service provider would only induce some L-type user to join the queue and hence reduce H-type users' welfare. Barring this exception, information design proves effective in improving welfare of both types over the no-information mechanism. Tying back to Proposition 3, we obtain that for all L-type users to choose their outside option under the no-information mechanism, the system must be already overwhelmed by H-type users, as captured by the condition $\lambda_H \geq \bar{\Lambda}_H$ on the arrival rate of H-type users.

5. Achieving First-Best

Having compared the effectiveness of information design against those of the two extreme signaling mechanisms, we now investigate its limitations. Specifically, in this section, we compare signaling mechanisms against Pareto-efficient admission policies and ask how limiting the requirement of ensuring obedience constraints is in terms of welfare improvement.

To better study this question, we consider the problem of maximizing the weighted welfare $W(\pi,\theta) = \theta W_L(\pi) + (1-\theta)W_H(\pi)$, both over the class of admission policies and the class of signaling mechanisms. As described in Section 2, each Pareto-efficient admission policy and signaling mechanism can be obtained as a maximizer of $W(\pi,\theta)$ for some $\theta \in [0,1]$. Furthermore, the specific θ for which a Pareto-efficient mechanism (or an admission policy) maximizes $W(\pi,\theta)$ captures the relative weight placed by the service provider in improving either type's welfare. For notational convenience, for any $\theta \in [0,1]$, we let $\mathsf{sm}(\theta)$ denote the signaling mechanism that maximizes $W(\pi,\theta)$ over $\pi \in \Pi_{SM}$, and $\mathsf{ap}(\theta)$ denote the admission policy[15] that does the same over Π_{AP}.

As we discussed in the conclusion of Section 3, Theorem 3 raises the possibility that there exists $\theta \in [0,1]$ such that $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$. The main point we make in this section is that, remarkably, for a wide range of θ, $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$; that is, the signaling mechanism $\mathsf{sm}(\theta)$ is Pareto-efficient within the class of admission policies. We have the following theorem, whose proof is given in Section EC.4 of the e-companion.

Theorem 6 (Achieving First-Best). *With $\bar{\Lambda}_H$ as defined in Proposition 3, the following holds.*

1. *For $\lambda_H \in [\bar{\Lambda}_H, 1]$, we have $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$ for all $\theta \in [0,1]$.*

2. *For $\lambda_H < \bar{\Lambda}_H$, there exists a $\theta(\lambda_L,\lambda_H) \in (0,1]$ such that for all $\theta > \theta(\lambda_L,\lambda_H)$ we have $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$, and for all $\theta < \theta(\lambda_L,\lambda_H)$ the signaling mechanism $\mathsf{sm}(\theta)$ is independent of θ.*

The preceding result has an interesting implication about the role of information design when signaling mechanisms achieve Pareto-dominance over Π_{AP}. In such cases, neither obedience constraint binds, since $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$. Thus, the obedience constraints impose no limitations on the service provider. In such cases, information design plays solely the role of a coordination device, directing some L-type users away from the queue and others to join the queue. In neither instance is the user indifferent between the recommended action and the alternative. This is unlike what happens in typical persuasion settings, where optimality requires indifference for at least some signals.

We also note that the two cases of the proposition are exactly the same as that of Theorem 5. In particular, when the no-information mechanism is Pareto-efficient, the set of Pareto-efficient signaling mechanisms is the same as the set of Pareto-efficient admission policies. Put differently, in instances where signaling mechanisms lack the power of admission policies, no-information is Pareto-dominated by some signaling mechanisms.

Finally, the equivalence of $\mathsf{sm}(\theta)$ and $\mathsf{ap}(\theta)$ is also appealing from an implementation point of view: the service provider can implement a signaling mechanism without the knowledge of user types. However, under an admission policy, the service provider observes the type of each arriving user and makes join and leave decisions on her behalf.

6. Numerical Analysis Under Linear Waiting Costs

To gain further comparative insights, in this section we augment our analytical results with a numerical analysis. We focus on the setting of linear utilities: $u_i(k) = V_i - c_i(k+1)$ for $i \in \{L,H\}$, where $V_i > 0$ denotes type-i users' value for the service, and $c_i > 0$ denotes the type-i users' waiting costs per unit time. (Note that the utility function $u_i(k)$ includes the waiting costs incurred due to time spent in the queue, as well as due to time spent receiving the service.) Since we focus on the notion of Pareto dominance, each users' utility can be scaled by an arbitrary positive number without any effect on our analysis. Thus, we normalize the utility functions by choosing the value $V_i = 1$ for each $i \in \{L,H\}$. With this normalization, we further assume that $c_L = c_H = c \in (0,1)$; this restricts our analysis to the setting where the two user types place the same relative weights on the value of service and the cost of waiting. Making this assumption of the homogeneity of the "inside option" enables us to neatly isolate the effects of the heterogeneity of the users' outside option.

Before we proceed with the analysis, we note that, in this setting, the quantity $\bar{\Lambda}_H$ in Proposition 3 is given

by $\bar{\Lambda}_{\mathsf{H}} = 1 - c \in (0,1)$. Thus, Proposition 3 implies that the no-information mechanism is Pareto-dominated for all $\lambda_{\mathsf{H}} \in [0, 1-c)$.

Next, we illustrate the qualitative insights of Proposition 3 and Theorems 4, 5, and 6 via numerical examples. First, in Figure 2, we plot the welfare of Pareto-efficient signaling mechanisms and admission policies for different values of $\lambda_{\mathsf{L}} \in \{0.13, 0.20, 0.30\}$, and $c = 0.15$. We fix $\lambda_{\mathsf{H}} = 1 - \lambda_{\mathsf{L}}$ in each case, to study the extreme setting where the service capacity exactly matches the total arrival rate $\lambda = \lambda_{\mathsf{L}} + \lambda_{\mathsf{H}} = 1$. For each value of λ_{L}, we also plot the full-information mechanism (fi) and the no-information mechanism (ni). First, observe that for $\lambda_{\mathsf{L}} = 0.13$, we have $\lambda_{\mathsf{H}} > \bar{\Lambda}_{\mathsf{H}} = 1 - c$, and, hence, from Proposition 3, the no-information mechanism (ni, green square) is Pareto-efficient. On the other hand, we note that the full-information mechanism (fi, green cross) is Pareto-dominated by a signaling mechanism (green star). Further, note that, as established in the first case of Theorem 6, $\mathsf{ap}(\theta) = \mathsf{sm}(\theta)$ for $\theta \in [0,1]$. On the other hand, for the other two values of λ_{L}, we see that the no-information mechanism achieves zero welfare for both types, and is Pareto-dominated in the class of signaling mechanisms. Additionally, even though the two Pareto frontiers do not coincide, they overlap considerably, particularly for $\lambda_{\mathsf{L}} = 0.20$. Finally, we observe that, as the proportion λ_{L} of users with viable outside option increases, the welfare of both user types increases.

We further complement the findings of Theorem 6 with numerical computations presented in Figure 3. Here, we plot the welfare of the signaling mechanism $\mathsf{sm}(\theta)$, the admission policy $\mathsf{ap}(\theta)$, and the full-information mechanism fi for $\theta \in \{0, 0.5, 1\}$, $c = 0.15$, and $\lambda = 1$. Note that $\theta = 0$ and $\theta = 1$ correspond to the extreme cases where the service provider seeks to maximize the welfare of one type, perhaps at the expense of the other. The case $\theta = 0.5$ corresponds to the case where the service provider values the two types equally. Together, these three cases provide a representative account of the service provider's potential objectives for welfare improvement. In these figures, in the region right of the green line, we have $\lambda_{\mathsf{H}} > \bar{\Lambda}_{\mathsf{H}} = 1 - c$. Thus, as shown in Theorem 6, we have $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$. For $\theta \in \{0.5, 1\}$, we see that, even for some values of $\lambda_{\mathsf{H}} < 1 - c$, the two are equal. Finally, we note that as $\lambda_{\mathsf{H}} \to 0$, we approach the homogeneous setting, and, as shown in Proposition 1, we observe the performance of the signaling mechanism $\mathsf{sm}(\theta)$ approaching that of the full-information mechanism in each case.

Finally, in Figure 4, we plot for each $c \in \{0.12, 0.24\}$, the values of $(\theta, \lambda_{\mathsf{H}})$ for which the Pareto-efficient admission policy $\mathsf{ap}(\theta)$ is the same as the Pareto-efficient signaling mechanism $\mathsf{sm}(\theta)$. (Here, again $\lambda_{\mathsf{L}} = 1 - \lambda_{\mathsf{H}}$.) In other words, for these values, information design plays mainly the role of a coordination device, inducing users to coordinate toward a better welfare outcome. In particular, neither obedience constraints bind for such values of $(\theta, \lambda_{\mathsf{H}})$. Observe that, as shown in Theorem 6, for any fixed λ_{H}, the values of θ for which this holds is an interval of the form $(\theta(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}}), 1]$. In particular, for $\lambda_{\mathsf{H}} > 1 - c$, this is the entire interval $[0,1]$. Conversely, for small-enough values of λ_{H}, that is, as we approach the homogeneous setting, we observe that this set is empty.

In Figure 4, the threshold corresponding to each point $(\theta, \lambda_{\mathsf{H}})$ is proportional to the color intensity used at the point. (Higher thresholds have lighter colors.) We observe that the threshold is nonzero for intermediate values of λ_{H} or sufficiently large θ. This highlights that it is possible to have $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$ while letting some L-type users join the queue. Further, note that for any fixed value of λ_{H}, as θ increases, the threshold value in the Pareto-efficient mechanism $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$ increases; as more weight is placed on L-type users' welfare, the Pareto-optimal signaling mechanism asks L-type users to join the queue for a larger range of queue-length values. (Also, as stated in Theorem 6, in the complement interval $[0, \theta(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}}))$ the threshold for $\mathsf{sm}(\theta)$ is independent of θ.) We also note that for any fixed θ, the values of λ_{H} for which $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$ is fairly complex, with it being a union of two intervals for some values of θ.

7. Extensions

In this section, we explain how our model can be generalized along two directions: (i) having H-type users with finite outside option (Section 7.1), and (ii) incorporating heterogeneity in service rates, along with a priority service discipline (Section 7.2). Additionally, in Sections EC.8 and EC.9 of the e-companion, we generalize our framework to allow for user abandonment and more than two types, respectively. Through analytical and numerical results, we illustrate that our key qualitative insights about the effectiveness of information design hold under all of the aforementioned generalizations.

7.1. Fully Persuadable Population

In our baseline model, we capture the level of need of H-type users by making the extreme assumption that H-type users have an outside option of $-\infty$. Consequently, such users cannot be not persuaded to avail the outside option in the model. Although such an assumption seems reasonable in certain contexts, such as the urgent care application discussed in Section 1.1, it is natural to examine the power of information design in settings where the entire user population is persuadable.

Toward that goal, we extend our model to incorporate finite outside options for both user types, with the H-type users having a worse outside option compared with the L-type users. In particular, normalizing the

L-type users' utility for the outside option to be 0, we denote the H-type users' utility for the outside option as $\ell_{\mathsf{H}} < 0$. Furthermore, for the ease of notation, we denote an H-type user's *incremental* utility for obtaining the service (over the outside option) as $u_{\mathsf{H}}(k) \triangleq u_{\mathsf{H}}^{O}(k) - \ell_{\mathsf{H}}$, where $u_{\mathsf{H}}^{O}(k)$ denotes the utility from availing the service when k users are already ahead in queue. In addition to Assumption 1 and paralleling the second condition therein, we assume that the difference $u_{\mathsf{H}}(n) - u_{\mathsf{H}}(n+1)$ is nonincreasing in n. Further, to gain structural insights, we make a *utility dominance* assumption, which requires that the utility of the H-type users dominates that of the L-type users at all queue lengths; that is, $u_{\mathsf{H}}(k) > u_{\mathsf{L}}(k)$ for all $k \geq 0$. (We observe that this dominance assumption is automatically satisfied in our baseline model with $\ell_{\mathsf{H}} = -\infty$.) In practice, a social service provider may not always be able to observe the type of a user. Moreover, ethical concerns may limit a service provider from making information provision depend on the users' outside options. Such limitations may make private signaling infeasible. Due to such considerations, we focus on *public signaling mechanisms*. Note that in our baseline model, public and private signaling are the same because high-need users always join irrespective of the belief.

Note that the utility dominance assumption implies that no (obedient) public signaling mechanism can provide a signal under which the L-type joins but the H-type does not join. Thus, we can restrict ourselves to signaling mechanisms that use three signals, denoted by $s \in \{0,1,2\}$, where signal $s = 0$ recommends no user type join, $s = 1$ recommends only the H-type joins, and $s = 2$ recommends both user types join. For any signaling mechanism, let $\pi_{k,j}$ denote the steady-state probability that the queue length is k and the signal sent is j. By analyzing the underlying birth-death chain of the queue, one can show that the steady-state distribution $\pi \triangleq \{\pi_{k,j} : k \geq 0, j = 0,1,2\}$ satisfies the following balance conditions:

$$\lambda\pi_{k,2} + \lambda_{\mathsf{H}}\pi_{k,1} = \sum_{j\in\{0,1,2\}} \pi_{k+1,j}, \quad \text{for } k \geq 0. \qquad \text{(BALANCE)}$$

Furthermore, for $i \in \{\mathsf{H},\mathsf{L}\}$ and $j \in \{0,1,2\}$, define $S_{i,j}(\pi) \triangleq \sum_{k=0}^{\infty} \pi_{k,j} u_i(k)$. Note that $S_{i,j}(\pi)$ is proportional to the expected utility obtained by a type i user for joining the queue upon receiving the signal j. Due to the utility dominance assumption, we have $S_{\mathsf{H},j}(\pi) \geq S_{\mathsf{L},j}(\pi)$. Thus, for π to be obedient, it must satisfy

$$S_{H,0}(\pi) \leq 0, \quad S_{\mathsf{L},1}(\pi) \leq 0 \leq S_{\mathsf{H},1}(\pi), \quad \text{and} \quad S_{\mathsf{H},2}(\pi) \geq 0. \qquad \text{(OBEDIENCE)}$$

With the above definitions, we have the welfare functions as

$$W_{\mathsf{L}}(\pi) = \lambda_{\mathsf{L}} S_{\mathsf{L},2}(\pi), \text{ and } W_{\mathsf{H}}(\pi) = \lambda_{\mathsf{H}}(S_{\mathsf{H},1}(\pi) + S_{\mathsf{H},2}(\pi)).$$

Following the same argument as described in Section 2, we can obtain the Pareto frontier of signaling mechanisms by solving the following linear program for each $\theta \in [0,1]$:

$$\begin{aligned} \max_{\pi} \quad & \theta W_{\mathsf{L}}(\pi) + (1-\theta) W_{\mathsf{H}}(\pi) \\ \text{subject to,} \quad & \text{(OBEDIENCE)}, \quad \text{(BALANCE)} \\ & \sum_{k=0}^{\infty}\sum_{j=0}^{2} \pi_{k,j} = 1, \quad \pi_{k,j} \geq 0 \\ & \qquad\qquad \text{for all } j = 0,1,2 \text{ and } k \geq 0. \end{aligned}$$

Finally, we define the class of threshold-signaling mechanisms as follows: for $0 \leq x \leq y$ with $x = m + a$ and $y = n + b$, where $m, n \in \mathbb{N}_0$ and $a, b \in [0,1)$, a threshold mechanism with thresholds x, y is given by

$$\begin{aligned} \pi_{k,2} &= Z\lambda^k \mathbf{I}\{0 \leq k < m\} + Z\lambda^m a \mathbf{I}\{k = m\}, \\ \pi_{k,1} &= Z\lambda^m (1-a)\mathbf{I}\{k = m\} + Z\lambda^m \lambda_a \lambda_{\mathsf{H}}^{k-m-1} \mathbf{I}\{m+1 \leq k \\ &\quad < n\} + Z\lambda^m \lambda_a \lambda_{\mathsf{H}}^{n-m-1} b \mathbf{I}\{k = n\} \\ \pi_{k,0} &= Z\lambda^m \lambda_a \lambda_{\mathsf{H}}^{n-m-1}(1-b)\mathbf{I}\{k = n\} \\ &\quad + Z\lambda^m \lambda_a \lambda_{\mathsf{H}}^{n-m} b \mathbf{I}\{k = n+1\}, \end{aligned}$$

where $\lambda_a \triangleq \lambda_{\mathsf{H}} + \lambda_{\mathsf{L}} a$ and $Z = Z(x,y)$ is a normalizing constant.[16] We denote the preceding mechanism as $\mathsf{Th}(x,y)$ and the corresponding welfare functions as $W_{\mathsf{H}}(x,y)$ and $W_{\mathsf{L}}(x,y)$.[17]

In the rest of this section, we first analytically establish the effectiveness of signaling mechanisms compared with the benchmarks of full information and no information (in Proposition 4). Then, we conduct a numerical analysis to illustrate the robustness of our insights (with respect to the H-type's outside option) and also discuss the structure of Pareto-efficient signaling mechanisms.

7.1.1. Mechanism Comparisons. The following proposition, which is in nature similar to Proposition 3, compares signaling mechanisms to the extreme cases of full- and no-information mechanisms. Characterization of these two mechanisms in this modified model naturally follows the ones specified in Section 2, and, for the sake of brevity, we do not repeat these definitions. For $i \in \{\mathsf{H},\mathsf{L}\}$, let m_i denote the smallest value of k for which $u_i(k) < 0$.

Proposition 4. *The following statements hold.*

1. *The full-information mechanism is Pareto-efficient within the class of signaling mechanisms if and only if it is Pareto-efficient within the class of admission policies. Furthermore, if* $W_{\mathsf{H}}(m_{\mathsf{L}}, m_{\mathsf{H}} - 1) \geq W_{\mathsf{H}}(m_{\mathsf{L}}, m_{\mathsf{H}})$ *or* $W_{\mathsf{L}}(m_{\mathsf{L}} - 1, m_{\mathsf{H}}) \geq W_{\mathsf{L}}(m_{\mathsf{L}}, m_{\mathsf{H}})$, *then the full-information mechanism is Pareto-dominated by a threshold-signaling mechanism.*

2. *The no-information mechanism is never Pareto-efficient in the class of admission policies. Furthermore, suppose that, under the no-information mechanism,* L*-type users join with positive probability. Then, the no-information mechanism is Pareto-dominated in the class of signaling mechanisms.*

The first part of Proposition 4 implies that if at least one of the types benefits from lowering its threshold below the full-information threshold, then there exists a signaling mechanism more effective than sharing full information. The second part implies that if without any information some L-type users still join (e.g., the system is not too overcrowded), then information design is more effective than sharing no information. Together, these two parts confirm that the power of information design persists in a population where all users are strategic and may decide not to join. The proof of Proposition 4 builds on the ideas used in the proof of Theorems 4 and 5 but also substantially departs from those proofs due to the difference in obedience constraints. The proof is presented in Section EC.5 of the e-companion.

7.1.2. Numerical Analysis. To further examine the effectiveness of signaling, we turn our attention to numerical analysis. We consider a setting analogous to that in Section 6 with arrival rates $\lambda_{\mathsf{L}} = 0.2$ and $\lambda_{\mathsf{H}} = 1 - \lambda_{\mathsf{L}}$. In particular, we assume that $u_{\mathsf{L}}(k) = 1 - c(k+1)$ and $u_{\mathsf{H}}(k) = 1 - c(k+1) - \ell_{\mathsf{H}}$, with $c = 0.13$ and $\ell_{\mathsf{H}} \in \{-1, -5, -10\}$.

In Figure 5, we plot the welfare of Pareto-efficient signaling mechanisms and admission policies, along with those of the full-information and the no-information mechanisms, for different values of the outside option ℓ_{H}. First, we observe that, in each setting, there exists a signaling mechanism that Pareto-dominates the full-information and the no-information mechanisms, thus demonstrating the effectiveness of information design in a fully persuadable user population. More importantly, we observe that, as the H-type users' outside options worsen, the Pareto frontier of admission policies approaches the Pareto frontier of the signaling mechanisms. Since the value of ℓ_{H} reflects the H-type users' need for the service, this numerical observation supports our broader conclusion that signaling is more effective when the user population is more heterogeneous in their need for service.

We conclude this discussion by noting that, in this generalized setting, the optimal signaling mechanism may not be a threshold mechanism. In Section EC.6 of the e-companion, we present two counterexamples that show that the optimal signaling can be "slightly" different from $\mathsf{Th}(x, y)$. However, our numerical analysis suggests that (1) such nonthreshold mechanisms only arise when $|\ell_{\mathsf{H}}|$ is relatively small, and (2) even in those cases, there exists a threshold mechanism that achieves a nearly identical welfare to that of the optimal signaling mechanism (See Figure EC.1 in Section EC.6 of the e-companion). Finally, we remark that, in the first part of Proposition 4, we provide sufficient conditions under which there exists a threshold mechanism that Pareto-dominates the full-information mechanism, implying the effectiveness of threshold mechanisms, even if they are not optimal.

7.2. Heterogeneity in Service Rates

Our baseline model assumes that the user types differ only in their utility (for both inside and outside options), but the service times across user types is homogeneous. However, in certain contexts, the heterogeneity in the need also translates to a heterogeneity in the service times. For instance, a critical patient arriving to an emergency room (ER) not only has a higher need for service but also requires longer service times, as compared with a noncritical patient. Given this practical concern, we consider a model where the user types not only differ in their need for service (through heterogeneity in their utility for service and outside option), but also in the service times—the service times of a type i user for $i \in \{\mathsf{H}, \mathsf{L}\}$ are assumed to be exponentially distributed with rate $\mu_i > 0$. To ensure stability, we assume that $\mu_i > \lambda_i$ for $i \in \{\mathsf{H}, \mathsf{L}\}$.

In such a setting, and especially when the user types are observable upon joining, it is reasonable to consider a preemptive priority service discipline, where any H-type user has priority over a L-type user. (For instance, such a discipline is natural in the ER setting mentioned above.) Under a preemptive priority service discipline, the state of the queue can be succinctly described as (m, n), where $m \geq 0$ is the number of H-type users and $n \geq 0$ is the number of L-type users in the queue. We let $u_i(m, n)$ denote the utility of a user of type $i \in \{\mathsf{H}, \mathsf{L}\}$ for joining the queue, and we assume that the outside option for the L-type user is 0, while that of the H-user is $-\infty$.[18]

Similar to our baseline model, to find the optimal signaling mechanisms in this setting, we formulate an infinite linear program, consisting of the obedience constraints and the balance conditions. Let $\pi^s_{m,n}$ denote the steady-state probability that the queue state is (m, n) and the signal sent to an arrival is $s \in \{0, 1\}$, where the signal $s = 1$ recommends that the L-type user join the queue, and $s = 0$ recommends that she choose the outside option. (Note that in this model, as in our baseline model, the H-type user always joins the queue.) Then, the balance conditions for the priority queue (with preemption) can be written as follows: for all $n \geq 0$,

$$
\begin{aligned}
&(\lambda_{\mathsf{H}} + \mu_{\mathsf{H}}) \sum_{s=0,1} \pi^s_{m,n} + \lambda_{\mathsf{L}} \pi^1_{m,n} \\
&= \mu_{\mathsf{H}} \sum_{s=0,1} \pi^s_{m+1,n} + \lambda_{\mathsf{H}} \sum_{s=0,1} \pi^s_{m-1,n} + \lambda_{\mathsf{L}} \pi^1_{m,n-1} \mathbf{I}\{n > 0\}, \quad \text{for } m \geq 1, \\
&(\lambda_{\mathsf{H}} + \mu_{\mathsf{L}} \mathbf{I}\{n > 0\}) \sum_{s=0,1} \pi^s_{0,n} + \lambda_{\mathsf{L}} \pi^1_{0,n} \\
&= \mu_{\mathsf{H}} \sum_{s=0,1} \pi^s_{1,n} + \mu_{\mathsf{L}} \sum_{s=0,1} \pi^s_{0,n+1} + \lambda_{\mathsf{L}} \pi_{0,n-1} \mathbf{I}\{n > 0\}, \quad \text{for } m = 0.
\end{aligned}
\tag{Pr-BAL}
$$

Given the steady-state distribution $\pi = \{\pi^s_{m,n} : m, n \geq 0, s = 0, 1\}$, the obedience constraints can be written as

$$\sum_{m,n\geq 0} \pi^1_{m,n} u_{\mathsf{L}}(m,n) \geq 0, \qquad \sum_{m,n\geq 0} \pi^0_{m,n} u_{\mathsf{L}}(m,n) \leq 0. \tag{Pr-OBD}$$

Here, the first constraint requires that the L-type user finds it optimal to join the queue upon receiving the signal $s = 1$, while the second constraints require that the user find it optimal to choose the outside option upon receiving the signal $s = 0$. Finally, the welfare of the two types can then be written as

$$W_{\mathsf{H}}(\pi) = \lambda_{\mathsf{H}} \sum_{m,n\geq 0} \sum_{s=0,1} \pi^s_{m,n} u_{\mathsf{H}}(m,n),$$
$$W_{\mathsf{L}}(\pi) = \lambda_{\mathsf{L}} \sum_{m,n\geq 0} \pi^1_{m,n} u_{\mathsf{L}}(m,n).$$

Note that since H-type users have (preemptive) priority over the L-type users, from their perspective, the queue only consists of H-type users. Consequently, the L-type users do not impose any externality on H-type users. Combined with the fact that H-type users always join, this implies that the welfare of H-type users is unaffected by the signaling mechanism. On the other hand, information design can still impact the welfare of L-type users, and, hence, a Pareto-efficient mechanism maximizes the welfare of L-type users and can be found by solving the following linear program:

$$\begin{aligned} \max_{\pi} \quad & W_{\mathsf{L}}(\pi) \\ \text{subject to,} \quad & \text{(Pr-OBD)}, \quad \text{(Pr-BAL)} \\ \sum_{m,n\geq 0} \quad & \sum_{s=0,1} \pi^s_{m,n} = 1, \quad \pi^s_{m,n} \geq 0, \\ & \text{for all } m, n \geq 0 \text{ and } s = 0, 1. \end{aligned}$$

For our numerical analysis of this model, we consider the case of linear waiting costs. From basic queueing analysis, we obtain $u_{\mathsf{H}}(m,n) = 1 - c_{\mathsf{H}}\left(\frac{m+1}{\mu_{\mathsf{H}}}\right)$ and

$$u_{\mathsf{L}}(m,n) = 1 - c_{\mathsf{L}}\left(\frac{m}{\mu_{\mathsf{H}} - \lambda_{\mathsf{H}}} + \frac{(n+1)\mu_{\mathsf{H}}}{\mu_{\mathsf{L}}(\mu_{\mathsf{H}} - \lambda_{\mathsf{H}})}\right).$$

In Figure 6, we plot the welfare of the Pareto-efficient signaling mechanisms (stars) and admission policies (circles) for $\mu_{\mathsf{H}} = 1, \mu_{\mathsf{L}} = 1.05, c_{\mathsf{H}} = c_{\mathsf{L}} = 0.15$, $\lambda_{\mathsf{H}} = 1 - \lambda_{\mathsf{L}}$, and for different values of $\lambda_{\mathsf{L}} \in \{0.13, 0.20, 0.30\}$. In each case, we also plot the full-information mechanism (fi, cross) and the no-information mechanism (ni, square). We observe that our qualitative insights continue to hold: when the population is sufficiently heterogeneous (e.g., when $\lambda_{\mathsf{L}} = 0.30$), the full-information and no-information mechanisms are Pareto-dominated by a signaling mechanism. This illustrates the power of information design over these benchmarks, even with heterogeneity in service times.

Furthermore, when $\lambda_{\mathsf{L}} = 0.30$ or 0.20, the Pareto-efficient signaling mechanism coincides with the Pareto-efficient admission policy. To highlight the power of signaling in achieving the first-best (i.e., the Pareto-efficient admission policy), in the right panel of Figure 6, we plot the welfare of L-type users under the Pareto-efficient signaling mechanism (sm) and the Pareto-efficient admission policy (ap) as λ_{L} varies from 0 to 1 (with $\lambda_{\mathsf{H}} = 1 - \lambda_{\mathsf{L}}$). We observe that when λ_{L} is low (i.e., the system is overcrowded by H-type users), then neither sm nor ap let any L-type user in, with both achieving welfare of 0. However, for moderate values of λ_{L}, sm and ap still coincide but now achieve positive welfare by letting some L-type users join. In Section EC.7 of the e-companion, we build on this numerical exercise to confirm that our qualitative findings remain the same under a wide range of service rates.

8. Conclusion

Social services often share two common features: they have limited capacity relative to their demand, and they aim to serve users with varied levels of needs. Reducing congestion for such services using price discrimination or admission control is often not feasible in this setting. However, the service provider can use its informational advantage, about service availability and wait times, to influence users' decisions in seeking a service by choosing what information to reveal. How effective will such a lever be? Our work seeks to answer this question. Adopting the framework of Bayesian persuasion, we study information design in a queueing system that serves users who are heterogeneous in their need for the service. We show that, by and large, information design provides a Pareto improvement in welfare of all user types when compared with simple mechanisms of sharing full information or no information. Further, we show that information design can go beyond and achieve the "first-best": it can achieve the same welfare outcomes as those of centralized admission policies that observe each user's type, and disregard user incentives. Finally, we show that our qualitative findings—on the benefits of well-designed information disclosure policies in the presence of need heterogeneity—continue to hold under various extensions to our model motivated from practical concerns. In sum, our results comprehensively exhibit a promising role for information design in improving welfare outcomes in congested social services.

Acknowledgments

The authors thank the department editor, the anonymous associate editor, and the reviewers for their constructive feedback. A preliminary version of this work appeared as an extended abstract in the proceedings of the 21st ACM Conference on Economics and Computation (EC'20), and we are grateful to the anonymous reviewers and conference participants for their valuable comments. The third

author gratefully acknowledges the Simons Institute for the Theory of Computing, as this work was done in part while the author attended the program on Online and Matching-Based Market Design.

Endnotes

[1] These websites provide wait time estimates for: JFK Medical Center, https://www.hackensackmeridianhealth.org/en/Locations/JFK-University-Medical-Center; and San Mateo Medical Center, http://204.114.51.208/. The wait time information for New York Public Housing can be found at: https://affordablehousingonline.com/open-section-8-waiting-lists/New-York. Applicants can check their wait times at this website: https://www.waitlistcheck.com/

[2] The quoted manager is Dr. Greg Rutledge, who is the chief of emergency medicine in one of Hamilton's regional hospitals (St. Joseph's Healthcare).

[3] In our base model, we focus on the first-come-first-serve queuing discipline. However, motivated by applications such as urgent care systems, we also study a preemptive priority discipline in Section 7.2 and demonstrate the robustness of our results under such a queuing discipline.

[4] We use the terms *Bayesian persuasion* and *information design* interchangeably.

[5] Since we focus on minimizing congestion, the service provider's preferences are aligned with the ex ante preferences of the users. We believe this preference alignment makes the social service provider more likely to keep its commitment.

[6] Further, in Sections EC.8 and EC.9 of the e-companion, we study two other extensions—respectively, exogenous abandonment and more than two types of users.

[7] As discussed in Section 1.1, some public housing authorities choose to disclose no information. Also, only some urgent care systems aim to provide real-time information that relies on (often) inaccurate estimations.

[8] Developing a model that captures information leakage while being grounded in practice as well as characterizing the optimal information design are interesting directions for future work.

[9] Although our contribution is theoretical, there is also an extensive literature on practical aspects of provision and prioritization of social services. For example, Brown and Watson (2018) examine the validity and reliability of a widely-used homelessness vulnerability assessment. Segall et al. (2016) outline criteria for kidney transplantation in elderly patients.

[10] This normalization of the service rate to 1 is without loss of generality.

[11] Here, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ denotes the set of nonnegative integers.

[12] The criterion of Pareto dominance evaluates signaling mechanisms at the ex ante stage, that is, upon a user arrival and before any information sharing or joining decisions. Such a criterion is natural when the focus is on assessing welfare outcomes in the aggregate, and in settings where the system state emerges endogenously from user behavior.

[13] We remark that, for the extreme cases of $\theta \in \{0,1\}$, not all maximizers of $W(\pi,\theta)$ over Π_{AP} (and Π_{SM}) may be Pareto-efficient. We discuss $\theta = 0$ here (the other extreme is similar). Since the optimization problem in this case puts no weight on the welfare of L-type users, there could be two maximizers π and π' such that $W_{\mathsf{H}}(\pi) = W_{\mathsf{H}}(\pi')$ but $W_{\mathsf{L}}(\pi) > W_{\mathsf{L}}(\pi')$. In that case, π' is Pareto-dominated by π and, hence, not Pareto-efficient. The recent work of Che et al. (2020) provides a refinement to address this issue: first optimize H-type's welfare and then among the optima, find the one that maximizes L-type's welfare. We follow the same refinement.

[14] The homogeneous setting where all users are H-type is uninteresting from the point of design, as all users join the queue regardless of their information due to no viable outside option.

[15] We note that, for certain values of θ, the maximizers of $W(\pi,\theta)$ over Π_{AP} (and Π_{SM}) may not be unique. To avoid burdensome notation, we let $\mathsf{ap}(\theta)$ (and $\mathsf{sm}(\theta)$) denote the set of all maximizers in such instances.

[16] Here, $Z(x,y)$ is given by $1/Z(x,y) = \sum_{k\le m}\lambda^k + \lambda^m\lambda_a\sum_{k=m+1}^{n}\lambda_{\mathsf{H}}^{k-m-1} + \lambda^m\lambda_a\lambda_{\mathsf{H}}^{n-m}b$.

[17] The welfare functions are given by $W_{\mathsf{L}}(x,y) = \lambda_{\mathsf{L}}Z(x,y)(\sum_{k<m}\lambda^k u_{\mathsf{L}}(k) + \lambda^m a u_{\mathsf{L}}(m))$ and $W_{\mathsf{H}}(x,y) = \lambda_{\mathsf{H}}Z(x,y)(\sum_{k=0}^{m}\lambda^k u_{\mathsf{H}}(k) + \lambda^m\lambda_a\sum_{k=m+1}^{n-1}\lambda_{\mathsf{H}}^{k-m-1}u_{\mathsf{H}}(k) + \lambda^m\lambda_a\lambda_{\mathsf{H}}^{n-m-1}b\cdot u_{\mathsf{H}}(n))$.

[18] In Section EC.7 of the e-companion, we discuss the above setting without the priority scheme.

References

Allon G, Bassamboo A, Gurvich I (2011) We will be right with you: Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59(6):1382–1394.

Ang E, Kwasnick S, Bayati M, Plambeck EL, Aratow M (2016) Accurate emergency department wait time prediction. *Manufacturing Service Oper. Management* 18(1):141–156.

Anunrojwong J, Iyer K, Lingenbrink D (2019) Persuading risk-conscious agents: A geometric approach. Preprint, submitted June 13, https://doi.org/10.2139/ssrn.3386273.

Arnosti N, Shi P (2020) Design of lotteries and wait-lists for affordable housing allocation. *Management Sci.* 66(6):2291–2307.

Ashlagi I, Jaillet P, Manshadi VH (2013) Kidney exchange in dynamic sparse heterogenous pools. Preprint, submitted January 15, https://doi.org/10.48550/arXiv.1301.3509.

Ashlagi I, Monachou F, Nikzad A (2020) Optimal dynamic allocation: Simplicity through information design. Preprint, submitted July 2, https://doi.org/10.2139/ssrn.3610386.

Ashlagi I, Burq M, Jaillet P, Manshadi V (2019) On matching and thickness in heterogeneous dynamic markets. *Oper. Res.* 67(4):927–949.

Baccara M, Lee S, Yariv L (2020) Optimal dynamic matching. *Theoret. Econom.* 15(3):1221–1278.

Balseiro SR, Gurkan H, Sun P (2019) Multiagent mechanism design without money. *Oper. Res.* 67(5):1417–1436.

Bergemann D, Morris S (2016) Bayes correlated equilibrium and the comparison of information structures in games. *Theoret. Econom.* 11(2):487–522.

Bimpikis K, Ehsani S, Mostagir M (2019) Designing dynamic contests. *Oper. Res.* 67(2):339–356.

Bimpikis K, Papanastasiou Y, Zhang W (2020) Information provision in two-sided platforms: Optimizing for supply. Preprint, submitted July 1, https://doi.org/10.2139/ssrn.3617351.

Brown M, Watson DP (2018) Reliability and validity of the vulnerability index-service prioritization decision assistance tool (VI-SPDAT) in real-world implementation. *J. Soc. Distress Homelessness* 27(2):110–117.

Candogan O, Drakopoulos K (2019) Optimal signaling of content accuracy: Engagement vs. misinformation. *Oper. Res.* 68(2):497–515.

Che YK, Tercieux O (2020) Optimal queue design. *Proc. 22nd ACM Conf. Econom. Comput.* (EC'21), 312–313. https://doi.org/10.1145/3465456.3467600.

Che YK, Kim J, Kojima F, Ryan CT (2020) Characterizing Pareto optima: Sequential utilitarian welfare maximization. Preprint, submitted August 25, https://doi.org/10.48550/arXiv.2008.10819.

Das S, Kamenica E, Mirka R (2017) Reducing congestion through information design. *Proc. 55th Annual Allerton Conf. Comm. Control Comput.* (IEEE, Piscataway, NJ), 1279–1284.

Doval L, Szentes B (2018) On the efficiency of queueing in dynamic matching markets. Working paper, http://www.laura-doval.com/on-the-efficiency-of-queueing-in-dynamic-matching-markets.

Drakopoulos K, Jain S, Randhawa RS (2018) Persuading customers to buy early: The value of personalized information provisioning. Preprint, submitted June 27, https://doi.org/10.2139/ssrn.3191629.

Gorokh A, Banerjee S, Iyer K (2021) From monetary to non-monetary mechanism design via artificial currencies. *Math. Oper. Res.* 46(3): 835–855.

Gross D, Shortle JF, Thompson JM, Harris CM (2018) *Fundamentals of Queueing Theory* (Wiley, New York).

Hassin R (2016) *Rational Queueing* (CRC Press, Boca Raton, FL).

Hassin R, Koshman A (2017) Profit maximization in the M/M/1 queue. *Oper. Res. Lett.* 45:436–441.

Ibrahim R (2018) Sharing delay information in service systems: A literature survey. *Queueing Systems* 89(1–2):49–79.

Kamenica E, Gentzkow M (2011) Bayesian persuasion. *Amer. Econom. Rev.* 101(6):2590–2615.

Kanoria Y, Saban D (2021) Facilitating the search for partners on matching platforms: Restricting agent actions. *Management Sci.* 67(10):5990–6029.

Kaplan EH (1984) Managing the demand for public housing. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Kremer I, Mansour Y, Perry M (2014) Implementing the "wisdom of the crowd." *J. Political Econom.* 122(5):988–1012.

Küçükgül C, Özer Ö, Wang S (2019) Engineering social learning: Information design of time-locked sales campaigns for online platforms. Preprint, submitted December 9, https://doi.org/10.2139/ssrn.3493744.

Leshno JD (2017) Dynamic matching in overloaded waiting lists. Preprint, submitted May 13, https://dx.doi.org/10.2139/ssrn.2967011.

Lingenbrink D, Iyer K (2018) Signaling in online retail: Efficacy of public signals. *Proc. 13th Workshop Econom. Networks Systems Comput.* (ACM, New York), 1, https://dl.acm.org/doi/10.1145/3230654.3230664.

Lingenbrink D, Iyer K (2019) Optimal signaling mechanisms in unobservable queues. *Oper. Res.* 67(5):1397–1416.

Mas-Colell A, Whinston MD, Green JR (1995) *Microeconomic Theory*, vol. 1 (Oxford University Press, New York).

Mitchell D (2020) Hamilton hospitals post real-time emergency room wait times online. GlobalNews.ca (January 9), https://globalnews.ca/news/6387106/hamilton-wait-times-online/.

Nahum Y, Sarne D, Das S, Shehory O (2015) Two-sided search with experts. *Autonomous Agents Multi-Agent Systems* 29(3): 364–401.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Papanastasiou Y, Bimpikis K, Savva N (2018) Crowdsourcing exploration. *Management Sci.* 64(4):1727–1746.

Prendergast C (2017) How food banks use markets to feed the poor. *J. Econom. Perspect.* 31(4):145–162.

Procaccia AD, Tennenholtz M (2009) Approximate mechanism design without money. *Proc. 10th ACM Conf. Electronic Commerce* (ACM, New York), 177–186.

Segall L, Nistor I, Pascual J, Mucsi I, Guirado L, Higgins R, Laecke SV, et al. (2016) Criteria for and appropriateness of renal transplantation in elderly patients with end-stage renal disease. *Transplantation* 100(10):55–65.

Xavier J (2017) Why hospital ER wait times are often wrong. *Insights* (August 7), https://www.gsb.stanford.edu/insights/why-hospital-er-wait-times-are-often-wrong.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Proofs and Auxiliary Results

In this e-companion, we provide the proofs of results in the main text, and discuss few extensions of our baseline model.

EC.1. Proofs from Section 3

In this section, we present the missing proofs from Section 3.

Proof of Lemma 2. The proof is immediate for $\lambda < 1$, since the sets Π_{AP} and Π_{SM} are compact (under the topology of weak convergence) and $W(\pi, \theta)$ is continuous in π for each $\theta \in [0,1]$. To see this, note that for any $\pi \in \Pi_{\mathsf{AP}}$, by Lemma EC.1 (stated at the end of this section), we have $\pi_k \leq \lambda^k \pi_0$ for all k. Hence, for $\lambda < 1$, Prohorov's theorem (Aliprantis and Border 2006) directly implies compactness of Π_{AP} and Π_{SM}.

Thus, for the rest of the proof, suppose $\lambda = 1$. Let $\Pi^{\mathsf{fi}}_{\mathsf{AP}}$ and $\Pi^{\mathsf{fi}}_{\mathsf{SM}}$ denote the set of admission policies and signaling mechanisms that are not Pareto-dominated by the full-information mechanism. We again use Prohorov's theorem to show that these sets are relatively compact, implying the existence of a maximizer of $W(\cdot, \theta)$ over the closures of these sets. The result then follows since any maximizer of $W(\pi, \theta)$ over the closure of $\Pi^{\mathsf{fi}}_{\mathsf{AP}}$ ($\Pi^{\mathsf{fi}}_{\mathsf{SM}}$) is also a maximizer over Π_{AP} (resp., Π_{SM}).

Thus, we first show that the set of distributions $\Pi^{\mathsf{fi}}_{\mathsf{AP}}$ is tight. Fix an $\epsilon > 0$. Let $W_{\mathsf{L}}(\mathsf{fi})$ and $W_{\mathsf{H}}(\mathsf{fi})$ denote the welfare of each type under the full information mechanism. Next, fix some large enough N to be chosen later. Consider a steady-state distribution $\pi \in \Pi_{\mathsf{AP}}$. We have the following expressions:

$$\begin{aligned} W_{\mathsf{H}}(\pi) &= \lambda_{\mathsf{H}} \sum_{n=0}^{\infty} \pi_n u_{\mathsf{H}}(n) \\ &\leq \lambda_{\mathsf{H}} u_{\mathsf{H}}(0) \left(\sum_{n<N} \pi_n \right) + \lambda_{\mathsf{H}} u_{\mathsf{H}}(N) \left(\sum_{n \geq N} \pi_n \right) \\ &\leq \lambda_{\mathsf{H}} u_{\mathsf{H}}(0) + \lambda_{\mathsf{H}} u_{\mathsf{H}}(N) \left(\sum_{n \geq N} \pi_n \right). \end{aligned}$$

where the first inequality follows from Assumption 1 and the second follows because $u_{\mathsf{H}}(0) > 0$. Similarly, we have for large enough $N > m_{\mathsf{fi}}$,

$$\begin{aligned} W_{\mathsf{L}}(\pi) &= \sum_{n=0}^{\infty} (\pi_{n+1} - \lambda_{\mathsf{H}} \pi_n) u_{\mathsf{L}}(n) \\ &\leq u_{\mathsf{L}}(0) \left(\sum_{n<N} (\pi_{n+1} - \lambda_{\mathsf{H}} \pi_n) \right) + u_{\mathsf{L}}(N) \left(\sum_{n \geq N} (\pi_{n+1} - \lambda_{\mathsf{H}} \pi_n) \right) \\ &= u_{\mathsf{L}}(0) \left(\pi_N - \pi_0 + (1 - \lambda_{\mathsf{H}}) \sum_{n=0}^{N-1} \pi_n \right) + u_{\mathsf{L}}(N) \left(-\lambda_{\mathsf{H}} \pi_N + (1 - \lambda_{\mathsf{H}}) \sum_{n>N} \pi_n \right) \end{aligned}$$

$$\leq (1-\lambda_{\mathsf{H}})u_{\mathsf{L}}(0) + u_{\mathsf{L}}(N)\left(-\pi_N + (1-\lambda_{\mathsf{H}})\sum_{n\geq N}\pi_n\right)$$
$$\leq \lambda_{\mathsf{L}}u_{\mathsf{L}}(0) + u_{\mathsf{L}}(N)\left(-\frac{1}{N+1} + \lambda_{\mathsf{L}}\sum_{n\geq N}\pi_n\right),$$

where the final inequality follows from (1) $\lambda_{\mathsf{L}} = 1-\lambda_{\mathsf{H}}$, (2) $u_{\mathsf{L}}(N) < 0$, and (3) $(N+1)\pi_N \leq \sum_{n=0}^{N}\pi_n \leq 1$ because of the detailed-balance conditions $\pi_n \leq \pi_{n-1}$. Thus, for $N \geq N_0^\epsilon = \frac{2}{\epsilon\lambda_{\mathsf{L}}}$, we obtain

$$W_{\mathsf{L}}(\pi) \leq \lambda_{\mathsf{L}}u_{\mathsf{L}}(0) + \lambda_{\mathsf{L}}u_{\mathsf{L}}(N)\left(-\frac{\epsilon}{2} + \sum_{n\geq N}\pi_n\right).$$

Since $\lim_{n\to\infty} u_i(n) = -\infty$, let N_1^ϵ be large enough so that $\max\{u_{\mathsf{L}}(k), u_{\mathsf{H}}(k)\} < -\frac{2}{\epsilon^2}$ for $k \geq N_1^\epsilon$. Then, for all $N \geq N^\epsilon = \max\{N_0^\epsilon, N_1^\epsilon\}$, we have

$$W_{\mathsf{H}}(\pi) \leq \lambda_{\mathsf{H}}\left(u_{\mathsf{H}}(0) - \frac{2}{\epsilon^2}\sum_{n\geq N}\pi_n\right)$$
$$W_{\mathsf{L}}(\pi) \leq \lambda_{\mathsf{L}}\left(u_{\mathsf{L}}(0) - \frac{2}{\epsilon^2}\left(-\frac{\epsilon}{2} + \sum_{n\geq N}\pi_n\right)\right).$$

Now, for any $\pi \in \Pi_{\mathsf{AP}}$, if $\sum_{n\geq N}\pi_n \geq \epsilon$, then we have $W_i(\pi) \leq \lambda_i u_i(0) - \frac{\lambda_i}{\epsilon}$ for $i \in \{\mathsf{L},\mathsf{H}\}$. For small enough $\epsilon > 0$, we obtain that $W_i(\pi) < W_i(\mathsf{fi})$ and hence π is Pareto dominated by the full-information mechanism, implying $\pi \notin \Pi_{\mathsf{AP}}^{\mathsf{fi}}$. Thus, we conclude that for all small enough $\epsilon > 0$, there exists an N^ϵ such that for all $\pi \in \Pi_{\mathsf{AP}}^{\mathsf{fi}}$, we have $\sum_{n\geq N^\epsilon}\pi_n < \epsilon$. Thus, the set of distributions $\Pi_{\mathsf{AP}}^{\mathsf{fi}}$ is tight.

Using Prohorov's theorem, we then conclude that $\Pi_{\mathsf{AP}}^{\mathsf{fi}}$ is relatively compact (under weak topology). The set $\Pi_{\mathsf{SM}}^{\mathsf{fi}}$, being a subset of $\Pi_{\mathsf{AP}}^{\mathsf{fi}}$, is also relatively compact. Since $W(\pi,\theta)$ is continuous in $\pi \in \Pi_{\mathsf{AP}}^{\mathsf{fi}}$ for any $\theta \in [0,1]$, we obtain that the maximizer of $W(\pi,\theta)$ over the closure of $\Pi_{\mathsf{AP}}^{\mathsf{fi}}$ (and, separately, $\Pi_{\mathsf{SM}}^{\mathsf{fi}}$) exists and is Pareto-efficient within Π_{AP} (resp., Π_{SM}). □

Proof of Theorem 1. Let $\pi \in \Pi_{\mathsf{AP}}$ be such that there exists an $m \geq 0$ with $\pi_{m+1} < \lambda\pi_m$ and $\lambda_{\mathsf{H}}\pi_{m+1} < \pi_{m+2}$. In words, this implies that under π, an arriving L-type user is asked to leave with positive probability if the queue length is m, and asked to join with positive probability if the queue length is $m+1$. We now show that such a π cannot be Pareto-efficient within Π_{SM}. We do this by constructing an $\hat{\pi} \in \Pi_{\mathsf{SM}}$ that Pareto-dominates π.

Towards that end, consider the following perturbation of π for small enough $\delta > 0$:

$$\hat{\pi}_k = \begin{cases} \pi_k & \text{if } k < m+1; \\ \pi_{m+1} + \delta\sum_{n>m+1}\pi_n & \text{if } k = m+1; \\ \pi_k(1-\delta) & \text{if } k > m+1. \end{cases}$$

First, it is straightforward to verify that $\hat{\pi}$ satisfies the detailed balance constraints in Lemma 1 for all small $\delta > 0$. In addition, we have

$$\begin{aligned} J(\hat{\pi}) &= \sum_{k=0}^{\infty} (\pi_{k+1} - \lambda_{\mathsf{H}} \pi_k) u_{\mathsf{L}}(k) + \delta \left(\sum_{k>m+1} \pi_k \right) (u_{\mathsf{L}}(m) - \lambda_{\mathsf{H}} u_{\mathsf{L}}(m+1)) \\ &\quad - \delta \pi_{m+2} u_{\mathsf{L}}(m+1) - \delta \sum_{k>m+1} (\pi_{k+1} - \lambda_{\mathsf{H}} \pi_k) u_{\mathsf{L}}(k) \\ &= J(\pi) + \delta \cdot \sum_{k>m+1} \pi_k \cdot (\, u_{\mathsf{L}}(m) - u_{\mathsf{L}}(k-1) - \lambda_{\mathsf{H}} (u_{\mathsf{L}}(m+1) - u_{\mathsf{L}}(k)) \,). \end{aligned}$$

Now, as $\lambda_{\mathsf{H}} < 1$, for any $k > m+1$, we have

$$\begin{aligned} u_{\mathsf{L}}(m) - u_{\mathsf{L}}(k-1) - \lambda_{\mathsf{H}} (u_{\mathsf{L}}(m+1) - u_{\mathsf{L}}(k)) &> u_{\mathsf{L}}(m) - u_{\mathsf{L}}(k-1) - u_{\mathsf{L}}(m+1) - u_{\mathsf{L}}(k) \\ &= (u_{\mathsf{L}}(m) - u_{\mathsf{L}}(m+1)) - (u_{\mathsf{L}}(k-1) - u_{\mathsf{L}}(k)) \\ &\geq 0, \end{aligned}$$

where we have used Assumption 1 in both inequalities. Specifically, the first inequality follows from the fact that $u_{\mathsf{L}}(k)$ is strictly decreasing in k and hence $u_{\mathsf{L}}(m+1) - u_{\mathsf{L}}(k) > 0$, and the second inequality follows from the fact that $u_{\mathsf{L}}(n) - u_{\mathsf{L}}(n+1)$ is non-increasing in n. Using this and the fact that $\pi_{m+2} > \lambda_{\mathsf{H}} \pi_{m+1} \geq 0$, we obtain that $J(\hat{\pi}) > J(\pi) \geq 0$. Hence the obedience constraint (JOIN) holds for $\hat{\pi}$.

By similar algebraic steps, we have

$$L(\hat{\pi}) = L(\pi) - \delta \cdot \sum_{k>m+1} \pi_k \cdot (\, u_{\mathsf{L}}(m) - u_{\mathsf{L}}(k-1) - \lambda (u_{\mathsf{L}}(m+1) - u_{\mathsf{L}}(k)) \,).$$

Using the fact that $\lambda \leq 1$, by a similar argument as before, we obtain that the parenthetical term is non-negative, and hence $L(\hat{\pi}) \leq L(\pi) \leq 0$. Thus, the obedience constraint (LEAVE) also holds for $\hat{\pi}$. Taken together, this implies we have $\hat{\pi} \in \Pi_{\mathsf{SM}}$.

Next, note that

$$\begin{aligned} W_{\mathsf{H}}(\hat{\pi}) &= \lambda_{\mathsf{H}} \sum_{n=0}^{\infty} \hat{\pi}_n u_{\mathsf{H}}(n) \\ &= W_{\mathsf{H}}(\pi) + \lambda_{\mathsf{H}} \delta \cdot \left(\sum_{k>m+1} \pi_k \cdot (u_{\mathsf{H}}(m+1) - u_{\mathsf{H}}(k)) \right). \end{aligned}$$

Since $u_{\mathsf{H}}(k)$ is strictly decreasing in k, we obtain $W_{\mathsf{H}}(\hat{\pi}) \geq W_{\mathsf{H}}(\pi)$. Finally, we have $W_{\mathsf{L}}(\hat{\pi}) = J(\hat{\pi}) > J(\pi) = W_{\mathsf{L}}(\pi)$. Thus, we obtain that $\hat{\pi}$ Pareto-dominates π.

From the above, we conclude that for any Pareto-efficient signaling mechanism $\pi \in \Pi_{\mathsf{SM}}$, it must be the case that whenever there exists an $m \geq 0$ with $\pi_{m+1} < \lambda \pi_m$, we have $\pi_{m+2} = \lambda_{\mathsf{H}} \pi_{m+1}$. This implies that π must have one of the following two structures:

1. for all $m \geq 0$, we have $\pi_{m+1} = \lambda \pi_m$; OR

2. there exists an $m \geq 0$ such that $\pi_{k+1} = \lambda \pi_k$ for $k < m$, $\pi_{m+1} < \lambda \pi_m$ and $\pi_{k+1} = \lambda_{\mathsf{H}} \pi_k$ for $k > m$.

In the first case, we have L-type users being asked to join the queue for all queue length, implying that π trivially has a threshold structure (with threshold equal to ∞). In the second case, the L-type users are asked to join with probability 1 for queue-lengths strictly less than m and asked to leave with probability 1 for queue-lengths strictly greater than m. Again, this implies a threshold structure for π, with threshold in the interval $[m, m+1]$.

Having shown the threshold structure of Pareto efficient signaling mechanisms, next we show that the corresponding threshold is less than or equal to the full-information threshold m_{fi}. Let $\pi \in \Pi_{\mathsf{SM}}$ have a threshold structure, with a threshold $x > m_{\mathsf{fi}}$, where $x = m + a$ with $m \in \mathbb{N}_0$ and $a \in [0, 1]$. Thus, we have $\pi_{k+1} = \lambda \pi_k$ for all $k < m$, and $\pi_{k+1} = \lambda_{\mathsf{H}} \pi_k$ for all $k > m$. Note, we allow $m = \infty$, which captures the case where $\pi_{k+1} = \lambda \pi_k$ for all $k \in \mathbb{N}_0$. Observe that $x > m_{\mathsf{fi}}$ implies that $m \geq m_{\mathsf{fi}}$, and hence the threshold structure of π implies $\pi_{m_{\mathsf{fi}}} > 0$.

We prove that such a distribution π cannot be Pareto efficient by constructing a $\hat{\pi} \in \Pi_{\mathsf{SM}}$ which Pareto dominates π. Consider $\hat{\pi}$ defined as follows:

$$\hat{\pi}_k = \begin{cases} \frac{1}{Z} \pi_k & \text{if } k \leq m_{\mathsf{fi}}; \\ \frac{1}{Z} \lambda_{\mathsf{H}}^{k - m_{\mathsf{fi}}} \pi_{m_{\mathsf{fi}}} & \text{if } k > m_{\mathsf{fi}}, \end{cases}$$

where $Z = \sum_{k \leq m_{\mathsf{fi}}} \pi_k + \pi_{m_{\mathsf{fi}}} \sum_{k > m_{\mathsf{fi}}} \lambda_{\mathsf{H}}^{k - m_{\mathsf{fi}}}$. Using the detailed balance constraints in Lemma 1, it follows that $\pi_k \geq \lambda_{\mathsf{H}}^{k - m_{\mathsf{fi}}} \pi_{m_{\mathsf{fi}}}$ for all $k > m_{\mathsf{fi}}$. Thus, as $\sum_k \pi_k = 1$, we have $Z \leq 1$.

Next, consider

$$\begin{aligned} J(\hat{\pi}) = \sum_{k=1}^{\infty} (\hat{\pi}_{k+1} - \lambda_{\mathsf{H}} \hat{\pi}_k) \, u_{\mathsf{L}}(k) &= \frac{1}{Z} \sum_{k < m_{\mathsf{fi}}} (\pi_{k+1} - \lambda_{\mathsf{H}} \pi_k) \, u_{\mathsf{L}}(k) \\ &> \frac{1}{Z} \sum_{k=1}^{\infty} (\pi_{k+1} - \lambda_{\mathsf{H}} \pi_k) \, u_{\mathsf{L}}(k) = \frac{1}{Z} \cdot J(\pi). \end{aligned}$$

Here, the inequality follows from the fact that $u_{\mathsf{L}}(k) < 0$ for $k \geq m_{\mathsf{fi}}$ and that $\pi_{m_{\mathsf{fi}}+1} - \lambda_{\mathsf{H}} \pi_{m_{\mathsf{fi}}} = \lambda_{\mathsf{L}} \min\{x - m_{\mathsf{fi}}, 1\} \pi_{m_{\mathsf{fi}}} > 0$. Since $J(\pi) \geq 0$ and $Z \leq 1$, we obtain that $J(\hat{\pi}) > J(\pi) \geq 0$. Hence, the obedience constraint (JOIN) holds for $\hat{\pi}$. Moreover, the threshold structure of π implies

$$L(\hat{\pi}) = \sum_{k=1}^{\infty} (\lambda \hat{\pi}_k - \hat{\pi}_{k+1}) \, u_{\mathsf{L}}(k) = \sum_{k=m_{\mathsf{fi}}}^{\infty} (\lambda \hat{\pi}_k - \hat{\pi}_{k+1}) \, u_{\mathsf{L}}(k) \leq 0.$$

Thus, the obedience constraint (LEAVE) also holds for $\hat{\pi}$. Hence, we obtain that $\hat{\pi} \in \Pi_{\mathsf{SM}}$.

Furthermore, for $\ell \leq m_{\mathsf{fi}}$, we have $\sum_{k \leq \ell} \hat{\pi}_k = \frac{1}{Z} \cdot \sum_{k \leq \ell} \pi_k$. Since, $Z \leq 1$, this implies $\sum_{k \leq \ell} \hat{\pi}_k \geq \sum_{k \leq \ell} \pi_k$ for all $\ell \leq m_{\mathsf{fi}}$. For $\ell > m_{\mathsf{fi}}$, after some algebra, we obtain

$$\sum_{k \leq \ell} \hat{\pi}_k - \sum_{k \leq \ell} \pi_k = \frac{1}{Z} \left(\sum_{q \leq m_{\mathsf{fi}}} \sum_{k > \ell} \pi_q \left(\pi_k - \pi_{m_{\mathsf{fi}}} \lambda_{\mathsf{H}}^{k - m_{\mathsf{fi}}} \right) + \sum_{q = m_{\mathsf{fi}}+1}^{\ell} \sum_{k > \ell} \pi_{m_{\mathsf{fi}}} \lambda_{\mathsf{H}}^{q - m_{\mathsf{fi}}} \left(\pi_k - \lambda_{\mathsf{H}}^{k-q} \pi_q \right) \right).$$

In Lemma EC.1 (stated at the end of this section), we show that $\pi_k \geq \pi_q \lambda_{\mathsf{H}}^{k-q}$ for all $k > q$. Thus, the right-hand side is non-negative, and hence, $\sum_{k\leq \ell} \hat{\pi}_k \geq \sum_{k\leq \ell} \pi_k$ for $\ell > m_{\mathsf{fi}}$ as well. Together, this implies that $\hat{\pi}$ is stochastically dominated by π. Since $u_{\mathsf{H}}(k)$ is strictly decreasing in k, we have

$$W_{\mathsf{H}}(\hat{\pi}) = \lambda_{\mathsf{H}} \sum_{k=0}^{\infty} \hat{\pi}_k u_{\mathsf{H}}(k) \geq \lambda_{\mathsf{H}} \sum_{k=0}^{\infty} \pi_k u_{\mathsf{H}}(k) = W_{\mathsf{H}}(\pi).$$

Finally, since $W_{\mathsf{L}}(\hat{\pi}) = J(\hat{\pi}) > J(\pi) = W_{\mathsf{L}}(\pi)$, we conclude that $\hat{\pi} \in \Pi_{\mathsf{SM}}$ Pareto-dominates π, and hence π cannot be Pareto-efficient within Π_{SM}. □

LEMMA EC.1. *For any $\pi \in \Pi_{\mathsf{AP}}$, and for any $k > q \in \mathbb{N}_0$, we have $\pi_k \geq \lambda_{\mathsf{H}}^{k-q} \pi_q$ and $\pi_k \leq \lambda^{k-q} \pi_q$. In particular, when $\lambda_{\mathsf{H}} > 0$, for any $\pi \in \Pi_{\mathsf{AP}}$, we have $\pi_k \in (0,1)$ for all $k \in \mathbb{N}_0$.*

Proof. The proof follows immediately from the detailed balance constraints in Lemma 1. □

Proof of Theorem 3. Since $\Pi_{\mathsf{SM}} \subset \Pi_{\mathsf{AP}}$, any signaling mechanism $\pi \in \Pi_{\mathsf{SM}}$ that is Pareto efficient within the class of admission policies must be so within the class of signaling mechanisms. Thus, it remains to show that for any signaling mechanism $\pi \in \Pi_{\mathsf{SM}}$ with $L(\pi) < 0$, if π is Pareto dominated by an admission policy, then it is Pareto dominated by a signaling mechanism.

By Theorem 1, we obtain that if π does not have a threshold structure, or if it has a threshold structure with threshold greater than the full-information threshold m_{fi}, then π is Pareto dominated within the class Π_{SM} of signaling mechanisms, and there is nothing to prove. Hence, suppose that π has a threshold structure with threshold smaller or equal to m_{fi}. This in turn implies that $J(\pi) > 0$, as a L-type user always receives non-negative utility upon joining the queue, and receives a positive utility if the queue is empty (which occurs with positive probability).

Since π is not Pareto-efficient within the class Π_{AP}, there exists an admission policy $\hat{\pi} \in \Pi_{\mathsf{AP}}$ that Pareto-dominates π. In particular, we have $W_i(\hat{\pi}) \geq W_i(\pi)$ for $i \in \{\mathsf{H}, \mathsf{L}\}$, with at least one inequality strict.

Next, let $\tilde{\pi} = (1-\epsilon)\pi + \epsilon\hat{\pi}$ for some $\epsilon \in (0,1]$ to be chosen later. By convexity of Π_{AP}, we have $\tilde{\pi} \in \Pi_{\mathsf{AP}}$. Furthermore, by linearity, we have $J(\tilde{\pi}) = (1-\epsilon)J(\pi) + \epsilon J(\hat{\pi})$ and $L(\tilde{\pi}) = (1-\epsilon)L(\pi) + \epsilon L(\hat{\pi})$. Since $J(\pi) > 0$ and $L(\pi) < 0$, for all small enough $\epsilon > 0$ we have $J(\tilde{\pi}) \geq 0$ and $L(\tilde{\pi}) \leq 0$. Thus, the obedience constraints (JOIN) and (LEAVE) hold for $\tilde{\pi}$, and hence $\tilde{\pi} \in \Pi_{\mathsf{SM}}$. Finally, again by linearity, we have

$$W_{\mathsf{L}}(\tilde{\pi}) = (1-\epsilon)W_{\mathsf{L}}(\pi) + \epsilon W_{\mathsf{L}}(\hat{\pi}) \geq W_{\mathsf{L}}(\pi)$$
$$W_{\mathsf{H}}(\tilde{\pi}) = (1-\epsilon)W_{\mathsf{H}}(\pi) + \epsilon W_{\mathsf{H}}(\hat{\pi}) \geq W_{\mathsf{H}}(\pi),$$

with at least one inequality strict. Thus, we obtain that the signaling mechanism $\tilde{\pi}$ Pareto-dominates π and hence π is not Pareto-efficient within the class Π_{SM}. □

EC.2. Structural Results

Before proceeding to present the missing proofs of Sections 4 and 5, we present three structural results. First, in Lemma EC.2, we characterize the equilibrium structure under the no-information mechanism. Then, in Lemma EC.3, we study the shape of welfare functions $W_{\mathsf{L}}(\cdot)$ and $W_{\mathsf{H}}(\cdot)$ for threshold mechanisms. Finally, in Lemma EC.4, we study the function $L(\cdot)$ defined in (LEAVE). We remark that the last two lemmas are used in the proofs of results in Sections 4 and 5.

LEMMA EC.2 **(Equilibrium structure under no-information mechanism)**. *For $p \in [0,1]$, let $\pi(p) \in \Pi_{\mathsf{AP}}$ be given by $\pi_n(p) = (1-\lambda_{\mathsf{L}}p-\lambda_{\mathsf{H}})(\lambda_{\mathsf{L}}p+\lambda_{\mathsf{H}})^n$. Then, the steady state distribution under the no-information mechanism* ni *is given by $\pi(p^{\mathsf{ni}}) \in \Pi_{\mathsf{SM}}$, for $p^{\mathsf{ni}} \in [0,1]$ that satisfies the following conditions:*

1. *if $\sum_{k=0}^{\infty} \lambda^k u_{\mathsf{L}}(k) \geq 0$ then $p^{\mathsf{ni}} = 1$;*
2. *if $\sum_{k=0}^{\infty} \lambda_{\mathsf{H}}^k u_{\mathsf{L}}(k) \leq 0$ then $p^{\mathsf{ni}} = 0$;*
3. *otherwise, $p^{\mathsf{ni}} \in (0,1)$ satisfies $\sum_{k=0}^{\infty} (\lambda_{\mathsf{L}}p^{\mathsf{ni}}+\lambda_{\mathsf{H}})^k u_{\mathsf{L}}(k) = 0$.*

Here, $p^{\mathsf{ni}} \in [0,1]$ denotes the probability under the no-information mechanism that a L*-type user joins the queue upon arrival.*

Proof of Lemma EC.2. First note that an arriving L-type user has no information about the queue length. Therefore, a symmetric equilibrium strategy consists of a probability p with which she joins the queue. Let $\pi_n(p)$ be the steady-state distribution corresponding to such a strategy. By detailed balance constraint, we have:

$$\pi_{n+1}(p) = (\lambda_{\mathsf{H}} + p\lambda_{\mathsf{L}})\pi_n(p),\ n \in \mathbb{N}_0$$

This implies $\pi_n(p) = (1-\lambda_{\mathsf{L}}p-\lambda_{\mathsf{H}})(\lambda_{\mathsf{L}}p+\lambda_{\mathsf{H}})^n$, $n \in \mathbb{N}_0$. A L-type users chooses p that maximizes her utility. This gives rise to the three cases listed in the statement of the lemma. □

LEMMA EC.3 **(Properties of welfare functions)**.

1. *The welfare function $W_{\mathsf{H}}(x)$ is strictly decreasing in $x \in \mathbb{R}_+$.*
2. *The welfare function $W_{\mathsf{L}}(x)$ is unimodal over $x \in \mathbb{R}_+$. Furthermore, $W_{\mathsf{L}}(x)$ is monotone between consecutive integers, initially increasing up to a maximum, and then decreasing.*
3. *The function $W(x,\theta) = \theta W_{\mathsf{L}}(x) + (1-\theta)W_{\mathsf{H}}(x)$ attains its maximum at an integer $m \leq m_{\mathsf{fi}}$.*

Proof of Lemma EC.3. The proof of the first statement follows from the fact the steady-state distribution under the threshold policy x is stochastically dominated by that under the threshold policy $\hat{x} > x$. Since $u_{\mathsf{H}}(k)$ is strictly decreasing in k, we thus obtain that $W_{\mathsf{H}}(x) > W_{\mathsf{H}}(\hat{x})$.

For the second statement, we show that (i) $W_{\mathsf{L}}(x)$ is monotone between consecutive integers, and (ii) $W_{\mathsf{L}}(x)$ is unimodal over integers, initially increasing up to a maximum, and then decreasing.

Together, theses two properties imply the unimodality of $W_{\mathsf{L}}(x)$ for $x \in \mathbb{R}_+$. Note that for $x = m+a$, where $m \in \mathbb{N}_0$ and $a \in [0,1)$, we have

$$W_{\mathsf{L}}(x) = \lambda_{\mathsf{L}} \cdot \frac{1}{\sum_{k=0}^{m} \lambda^k + \frac{\lambda^m(\lambda_{\mathsf{H}} + a\lambda_{\mathsf{L}})}{1-\lambda_{\mathsf{H}}}} \cdot \left(\sum_{k=0}^{m-1} \lambda^k u_{\mathsf{L}}(k) + a\lambda^m u_{\mathsf{L}}(m) \right).$$

Since this is of the form $\frac{\alpha+\beta a}{\gamma+\delta a}$, where $\alpha, \beta, \gamma, \delta$ are independent of a, we obtain that $W_{\mathsf{L}}(m+a)$ is monotone in $a \in [0,1)$. Thus, we conclude that $W_{\mathsf{L}}(x)$ is monotone between consecutive integers. It is straightforward to verify that $W_{\mathsf{L}}(x)$ is continuous, and hence the maximum of $W_{\mathsf{L}}(x)$ is attained at an integer.

For $m \in \mathbb{N}_0$, we have

$$W_{\mathsf{L}}(m) = \lambda_{\mathsf{L}} \cdot \frac{1}{\sum_{k=0}^{m} \lambda^k + \frac{\lambda^m \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}}} \cdot \sum_{k=0}^{m-1} \lambda^k u_{\mathsf{L}}(k) = \lambda_{\mathsf{L}} \cdot \Gamma(m) \cdot \Lambda(m) = \lambda_{\mathsf{L}} \Phi(m),$$

where $\Gamma(m) = \frac{1}{\sum_{k=0}^{m} \lambda^k + \frac{\lambda^m \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}}}$, $\Lambda(m) = \sum_{k=0}^{m-1} \lambda^k u_{\mathsf{L}}(k)$, and $\Phi(m) = \Gamma(m)\Lambda(m)$.

In the following, we show Φ is unimodal by establishing that if Φ decreases at some integer $m \in \mathbb{N}_0$, then it decreases at all integers $k \geq m$. Towards that goal, for any function $f : \mathbb{N}_0 \to \mathbb{R}$, let $\Delta f(m) \triangleq f(m) - f(m-1)$ denote the finite difference at m. Then, we have

$$\begin{aligned}
\Delta\Gamma(m) &= \frac{1}{\sum_{k=0}^{m} \lambda^k + \frac{\lambda^m \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}}} - \frac{1}{\sum_{k=0}^{m-1} \lambda^k + \frac{\lambda^{m-1} \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}}} = -\lambda^{m-1} \left(\frac{\lambda - \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}} \right) \Gamma(m)\Gamma(m-1), \\
\Delta\Lambda(m) &= \lambda^{m-1} u_{\mathsf{L}}(m-1),
\end{aligned}$$

and hence,

$$\begin{aligned}
\Delta\Phi(m) &= \Lambda(m)\Delta\Gamma(m) + \Gamma(m-1)\Delta\Lambda(m) \\
&= -\lambda^{m-1} \left(\frac{\lambda - \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}} \right) \Gamma(m)\Gamma(m-1)\Lambda(m) + \lambda^{m-1} u_{\mathsf{L}}(m-1)\Gamma(m-1) \\
&= \lambda^{m-1}\Gamma(m-1) \left(u_{\mathsf{L}}(m-1) - \left(\frac{\lambda - \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}} \right) \Phi(m) \right). \qquad \text{(EC.1)}
\end{aligned}$$

Substituting $\Phi(m) = \Phi(m-1) + \Delta\Phi(m)$ into (EC.1) and after some algebra, we obtain for all $k \in \mathbb{N}_0$,

$$\left(1 + \left(\frac{\lambda - \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}} \right) \lambda^{k-1}\Gamma(k-1) \right) \Delta\Phi(k) = \lambda^{k-1}\Gamma(k-1) \left(u_{\mathsf{L}}(k-1) - \left(\frac{\lambda - \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}} \right) \Phi(k-1) \right). \quad \text{(EC.2)}$$

Now, suppose $\Delta\Phi(m) = \Phi(m) - \Phi(m-1) \leq 0$ for some $m \geq 1$. Since $\Gamma(m-1)$ is positive, from (EC.1) we obtain $u_{\mathsf{L}}(m-1) \leq \left(\frac{\lambda - \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}} \right) \Phi(m)$. Using the expression (EC.2) with $k = m+1$, we get

$$\begin{aligned}
\left(1 + \left(\frac{\lambda - \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}} \right) \lambda^m \Gamma(m) \right) \Delta\Phi(m+1) &= \lambda^m \Gamma(m) \left(u_{\mathsf{L}}(m) - \left(\frac{\lambda - \lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}} \right) \Phi(m) \right) \\
&\leq \lambda^m \Gamma(m) \left(u_{\mathsf{L}}(m) - u_{\mathsf{L}}(m-1) \right) \\
&\leq 0,
\end{aligned}$$

where we have used $u_{\mathsf{L}}(m-1) \le \left(\frac{\lambda-\lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}}\right)\Phi(m)$ in the first inequality, and the fact that $u_{\mathsf{L}}(\cdot)$ is decreasing in the second inequality. This in turn implies that $\Delta\Phi(m+1) \le 0$.

Thus, by induction, we obtain that if $\Delta\Phi(m) \le 0$ then $\Delta\Phi(m+k) \le 0$ for all $k \ge 0$. This proves the unimodality of Φ and hence that of $W_{\mathsf{L}} = \lambda_{\mathsf{L}}\Phi$. Finally, note that $W_{\mathsf{L}}(1) - W_{\mathsf{L}}(0) = \lambda_{\mathsf{L}}(1-\lambda_{\mathsf{H}})\frac{u_{\mathsf{L}}(0)}{1-\lambda_{\mathsf{H}}+\lambda} > 0$. Thus, $W_{\mathsf{L}}(m)$ initially increases up to a maximum, and then decreases subsequently.

For the third statement, note that for $x = m+a$, where $m \in \mathbb{N}_0$ and $a \in [0,1)$, we have

$$\begin{aligned} W(x,\theta) &= \theta W_{\mathsf{L}}(x) + (1-\theta)W_{\mathsf{H}}(x) \\ &= \frac{1}{\sum_{k=0}^{m}\lambda^k + \frac{\lambda^m(\lambda_{\mathsf{H}}+a\lambda_{\mathsf{L}})}{1-\lambda_{\mathsf{H}}}} \left(\theta\lambda_{\mathsf{L}}\left(\sum_{k=0}^{m-1}\lambda^k u_{\mathsf{L}}(k) + a\lambda^m u_{\mathsf{L}}(m)\right)\right. \\ &\qquad \left. +(1-\theta)\left(\lambda_{\mathsf{H}}\sum_{k=0}^{m}\lambda^k u_{\mathsf{H}}(k) + (a\lambda_{\mathsf{L}}+\lambda_{\mathsf{H}})\sum_{k>m}\lambda^m\lambda_{\mathsf{H}}^{k-1-m}u_{\mathsf{H}}(k)\right)\right). \end{aligned}$$

Again, this is of the form $\frac{\alpha+\beta a}{\gamma+\delta a}$, where $\alpha,\beta,\gamma,\delta$ are independent of $a \in [0,1)$. Thus, we obtain that $W(m+a,\theta)$ is monotone in $a \in [0,1)$, and hence $W(x,\theta)$ is monotone between consecutive integers. Since $W(x,\theta)$ is continuous in x, the maximum of $W(x,\theta)$ is attained at an integer. □

Lemma EC.4 **(Properties of the** (LEAVE) **function)**. *For $x \in \mathbb{R}_+$, the function $L(x)$ is strictly decreasing as long as it is non-negative, subsequent to which it stays negative. Formally, we have $L(x) \le \max\{\inf_{0\le u\le x} L(u), 0\}$.*

Proof of Lemma EC.4. Consider a threshold policy $x = m+a$, where $m \in \mathbb{N}_0$ and $a \in [0,1)$. We have

$$\begin{aligned} L(x) &= \sum_{k=0}^{\infty}(\lambda\pi_k - \pi_{k+1})u_{\mathsf{L}}(k) \\ &= \lambda_{\mathsf{L}}\pi_m(1-a)u_{\mathsf{L}}(m) + \lambda_{\mathsf{L}}\sum_{k=1}^{\infty}\pi_{m+k}u_{\mathsf{L}}(m+k) \\ &= \lambda_{\mathsf{L}}\cdot\frac{\lambda^m(1-a)u_{\mathsf{L}}(m) + \lambda^m(\lambda_{\mathsf{H}}+\lambda_{\mathsf{L}}a)\sum_{k=1}^{\infty}\lambda_{\mathsf{H}}^{k-1}u_{\mathsf{L}}(m+k)}{\sum_{k=0}^{m}\lambda^k + \frac{\lambda^m(\lambda_{\mathsf{H}}+a\lambda_{\mathsf{L}})}{1-\lambda_{\mathsf{H}}}}. \end{aligned}$$

Since this is a ratio of two linear functions of a, we obtain that it is monotone in a, and hence, it suffices to analyze $L(x)$ as a function over integers. After some algebra, we have

$$\begin{aligned} \frac{1}{\lambda_{\mathsf{L}}}L(m) &= \frac{\lambda^m\sum_{k=0}^{\infty}\lambda_{\mathsf{H}}^k u_{\mathsf{L}}(m+k)}{\sum_{k=0}^{m}\lambda^k + \frac{\lambda^m\lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}}} \\ &= \frac{\lambda^m\sum_{k=0}^{\infty}\lambda_{\mathsf{H}}^k}{\sum_{k=0}^{m}\lambda^k + \frac{\lambda^m\lambda_{\mathsf{H}}}{1-\lambda_{\mathsf{H}}}}\cdot\frac{\sum_{k=0}^{\infty}\lambda_{\mathsf{H}}^k u_{\mathsf{L}}(m+k)}{\sum_{k=0}^{\infty}\lambda_{\mathsf{H}}^k}. \end{aligned}$$

Now, both factors on the right-hand side are strictly decreasing in m. Further, the first factor is positive. If $L(m)$ is non-negative, then the second factor is non-negative, and hence $L(m+1) -$

$L(m) < 0$. On the other hand, if $L(m) < 0$, then the second factor is negative, and since it is decreasing, we obtain $L(m+1) < 0$ as well. Thus, we conclude that $L(x)$ is strictly decreasing as long as it is non-negative, subsequent to which it stays negative. Formally, we have $L(x) \leq \max\{\inf_{0\leq u\leq x} L(y), 0\}$. $\square$

EC.3. Proofs from Section 4

Proof of Proposition 1. First note that $0 < W_{\mathsf{L}}(\mathsf{fi}) \leq W_{\mathsf{L}}(\mathsf{sm})$ simply follows from the observation that fi is a feasible signaling mechanism. Thus its welfare is a lower bound on that achieved by the optimal signaling mechanism sm.

Next, we prove $W_{\mathsf{L}}(\mathsf{fi}) \geq \beta_{\mathsf{fi}} W_{\mathsf{L}}(\mathsf{sm})$. The proof consists of two steps. In the first step, we show $x_{\mathsf{sm}} \geq m_{\mathsf{fi}} - 1$, where $x_{\mathsf{sm}} \in \mathbb{R}_+$ is the threshold of the sm mechanism. We prove this by showing that the second obedience constraint, (LEAVE), will not be satisfied if the threshold is below $m_{\mathsf{fi}} - 1$. More precisely, let π^{sm} denote the steady-state distribution corresponding to sm mechanism, and let $x_{\mathsf{sm}} = m + a$ where $m \in \mathbb{N}_0$ and $a \in [0,1)$. Then

$$L(\pi^{\mathsf{sm}}) = \begin{cases} \lambda_{\mathsf{L}} \pi_m (1-a) \cdot u_{\mathsf{L}}(m) + \lambda_{\mathsf{L}} \pi_{m+1} \cdot u_{\mathsf{L}}(m+1), & \text{if } a > 0; \\ \lambda_{\mathsf{L}} \pi_m \cdot u_{\mathsf{L}}(m), & \text{if } a = 0. \end{cases}$$

Here, the first case follows from the fact that, under the optimal signaling mechanism sm, a user is asked to leave with probability $1-a$ if the queue length equals m, which occurs with probability π_m, and asked to leave with probability 1 if the queue-length equals $m+1$, which occurs with probability π_{m+1}. The second case follows analogously.

Since $\pi^{\mathsf{sm}} \in \Pi_{\mathsf{SM}}$, we have $L(\pi^{\mathsf{sm}}) \leq 0$. This condition, along with the fact that $u_{\mathsf{L}}(\cdot)$ is strictly decreasing, forces $u_{\mathsf{L}}(m+1) < 0$ if $a > 0$, and $u_{\mathsf{L}}(m) \leq 0$ and $u_{\mathsf{L}}(m+1) < 0$ if $a = 0$. In both cases, we have $m + 1 \geq m_{\mathsf{fi}}$, and hence $x_{\mathsf{sm}} = m + a \geq m_{\mathsf{fi}} - 1 + a \geq m_{\mathsf{fi}} - 1$. Further, from Theorem 1, we have $x_{\mathsf{sm}} \leq m_{\mathsf{fi}}$. Putting these two together, we obtain

$$x_{\mathsf{sm}} \in [m_{\mathsf{fi}} - 1, m_{\mathsf{fi}}].$$

Since $W_{\mathsf{L}}(x)$ is monotone between integers (as established in Lemma EC.3), we thus obtain that $W_{\mathsf{L}}(\mathsf{fi}) = W_{\mathsf{L}}(\mathsf{sm})$ if and only if $W_{\mathsf{L}}(m_{\mathsf{fi}}) \geq W_{\mathsf{L}}(m_{\mathsf{fi}} - 1)$. Furthermore, we have $W_{\mathsf{L}}(\mathsf{sm}) \leq \max\{W_{\mathsf{L}}(m_{\mathsf{fi}} - 1), W_{\mathsf{L}}(m_{\mathsf{fi}})\}$. Now,

$$\begin{aligned} W_{\mathsf{L}}(m_{\mathsf{fi}} - 1) &= \frac{\sum_{n=0}^{m_{\mathsf{fi}}-2} \lambda_{\mathsf{L}}^n u_{\mathsf{L}}(n)}{\sum_{n=0}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{L}}^n} \\ &\leq \frac{\sum_{n=0}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{L}}^n u_{\mathsf{L}}(n)}{\sum_{n=0}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{L}}^n} = \left(\frac{\sum_{n=0}^{m_{\mathsf{fi}}} \lambda_{\mathsf{L}}^n}{\sum_{n=0}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{L}}^n} \right) \cdot \frac{\sum_{n=0}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{L}}^n u_{\mathsf{L}}(n)}{\sum_{n=0}^{m_{\mathsf{fi}}} \lambda_{\mathsf{L}}^n} = \frac{1}{\beta_{\mathsf{fi}}} \cdot W_{\mathsf{L}}(\mathsf{fi}). \end{aligned}$$

Here, in the inequality follows from the fact that $u_{\mathsf{L}}(m_{\mathsf{fi}} - 1) \geq 0$, and the first and the second equalities follow from the definition of a threshold mechanism. In the final equality, we have used

the definition of β_{fi}. Thus, taken together, we obtain $W_{\mathsf{L}}(\mathsf{fi}) \geq \beta_{\mathsf{fi}} W_{\mathsf{L}}(\mathsf{sm})$. The statement of the proposition follows after noting that $\beta_{\mathsf{fi}} \geq 1 - \frac{1}{m_{\mathsf{fi}}+1}$ for all $\lambda_{\mathsf{L}} \leq 1$. $\square$

Proof of Proposition 2. Recall that p^{ni} denotes the probability with which a L-type user joins under the no-information mechanism. Since $\lambda_{\mathsf{H}} = 0$, we note that $p^{\mathsf{ni}} > 0$, as a L-type user will find it optimal to join the queue if no other such user does so. Consequently, we consider the cases $p^{\mathsf{ni}} = 1$ and $p^{\mathsf{ni}} \in (0,1)$. Suppose $p^{\mathsf{ni}} = 1$. Then we have

$$\begin{aligned}
W_{\mathsf{L}}(\mathsf{ni}) &= \sum_{n \in \mathbb{N}_0} (1-\lambda_{\mathsf{L}}) \lambda_{\mathsf{L}}^n u_{\mathsf{L}}(n) \\
&\leq \sum_{n < m_{\mathsf{fi}}} (1-\lambda_{\mathsf{L}}) \lambda_{\mathsf{L}}^n u_{\mathsf{L}}(n) + u_{\mathsf{L}}(m_{\mathsf{fi}}) \lambda_{\mathsf{L}}^{m_{\mathsf{fi}}} \\
&= \left(\sum_{n=0}^{m_{\mathsf{fi}}} (1-\lambda_{\mathsf{L}}) \lambda_{\mathsf{L}}^n \right) \cdot \frac{\sum_{n=0}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{L}}^n u_{\mathsf{L}}(n)}{\sum_{n=0}^{m_{\mathsf{fi}}} \lambda_{\mathsf{L}}^n} + u_{\mathsf{L}}(m_{\mathsf{fi}}) \lambda_{\mathsf{L}}^{m_{\mathsf{fi}}} \\
&= \left(1 - \lambda_{\mathsf{L}}^{m_{\mathsf{fi}}+1}\right) \cdot W_{\mathsf{L}}(\mathsf{fi}) + u_{\mathsf{L}}(m_{\mathsf{fi}}) \lambda_{\mathsf{L}}^{m_{\mathsf{fi}}} \\
&< \left(1 - \lambda_{\mathsf{L}}^{m_{\mathsf{fi}}+1}\right).
\end{aligned}$$

Here, we use the fact that $u_{\mathsf{L}}(k)$ is decreasing k in the first inequality, and the final inequality follows from $u_{\mathsf{L}}(m_{\mathsf{fi}}) < 0$. On the other hand, if $p^{\mathsf{ni}} \in (0,1)$, we have $W_{\mathsf{L}}(\mathsf{ni}) = 0 < (1 - \lambda_{\mathsf{L}}^{m_{\mathsf{fi}}+1}) W_{\mathsf{L}}(\mathsf{fi})$. $\square$

Proof of Proposition 3. Recall from Theorem 4 that the full-information mechanism is Pareto-efficient if and only if $W_{\mathsf{L}}(m_{\mathsf{fi}}) - W_{\mathsf{L}}(m_{\mathsf{fi}} - 1) > 0$. By a little algebra, this condition can be shown to be equivalent to

$$f(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}}) \triangleq \lambda_{\mathsf{L}} u_{\mathsf{L}}(0) - \lambda_{\mathsf{L}} \sum_{k=1}^{m_{\mathsf{fi}}-1} (\lambda_{\mathsf{H}} + \lambda_{\mathsf{L}})^k (u_{\mathsf{L}}(k-1) - u_{\mathsf{L}}(k)) - (1-\lambda_{\mathsf{H}}) u_{\mathsf{L}}(m_{\mathsf{fi}} - 1) < 0.$$

It is straightforward to verify that $f(0, \lambda_{\mathsf{H}}) < 0$, $f(1-\lambda_{\mathsf{H}}, \lambda_{\mathsf{H}}) = 0$, $\partial_{\mathsf{L}} f(0, \lambda_{\mathsf{H}}) = u_{\mathsf{L}}(0) - \sum_{k=1}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{H}}^k (u_{\mathsf{L}}(k-1) - u_{\mathsf{L}}(k)) \geq u_{\mathsf{L}}(m_{\mathsf{fi}} - 1) > 0$, and $\partial_{\mathsf{L}}^2 f < 0$ for $\lambda_{\mathsf{L}} \in [0, 1-\lambda_{\mathsf{H}}]$, where ∂_{L} denotes the partial derivative with respect to λ_{L}. These facts imply that for any fixed $\lambda_{\mathsf{H}} \in (0,1)$, the function $f(\cdot, \lambda_{\mathsf{H}})$ has a root $\bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}}) \in (0, 1-\lambda_{\mathsf{H}}]$ satisfying $f(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}}) < 0$ for $\lambda_{\mathsf{L}} < \bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}})$ and $f(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}}) > 0$ for $\bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}}) < \lambda_{\mathsf{L}} < 1 - \lambda_{\mathsf{H}}$. Thus, we obtain that the full-information mechanism is Pareto-efficient if and only if $\lambda_{\mathsf{L}} < \bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}})$. Finally, the definition of f, along with some straightforward algebra, yields the following lower-bound:

$$\bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}}) \geq \frac{u_{\mathsf{L}}(m_{\mathsf{fi}} - 1)}{u_{\mathsf{L}}(0) - \sum_{k=1}^{m_{\mathsf{fi}}-1} \lambda_{\mathsf{H}}^k (u_{\mathsf{L}}(k-1) - u_{\mathsf{L}}(k))} \cdot (1 - \lambda_{\mathsf{H}}).$$

In order to prove the second part, we note that the proof of the first part implies that if $\lambda_{\mathsf{L}} < \bar{\Lambda}_{\mathsf{L}}(\lambda_{\mathsf{H}})$, we have $J(m_{\mathsf{fi}} - 1) = W_{\mathsf{L}}(m_{\mathsf{fi}} - 1) \geq W_{\mathsf{L}}(m_{\mathsf{fi}}) > 0$. Furthermore, the assumption $L(m_{\mathsf{fi}} - 1) \leq 0$ implies that the threshold mechanism with threshold of $m_{\mathsf{fi}} - 1$ is an obedient mechanism. This further implies that the efficient signaling mechanism has a threshold of at most $m_{\mathsf{fi}} - 1$. To see why,

suppose $x_{\mathsf{sm}} > m_{\mathsf{fi}} - 1$. As established in Lemma EC.3, $W_{\mathsf{H}}(m_{\mathsf{fi}}-1) > W_{\mathsf{H}}(x_{\mathsf{sm}})$ and $W_{\mathsf{L}}(m_{\mathsf{fi}}-1) \geq W_{\mathsf{L}}(x_{\mathsf{sm}})$ implying that the threshold mechanism with threshold $m_{\mathsf{fi}}-1$ Pareto dominates the one with threshold x_{sm} which is a contradiction.

In light of the above observations, we have $W_i(\mathsf{sm}) \geq W_i(m_{\mathsf{fi}}-1)$, for $i \in \{\mathsf{L},\mathsf{H}\}$. Thus in the following, we establish a multiplicative gap between $W_i(m_{\mathsf{fi}}-1)$ and $W_i(m_{\mathsf{fi}})$ for each $i \in \{\mathsf{L},\mathsf{H}\}$.

We start with the L-type users. Observe that, for any threshold mechanism with threshold $m \leq m_{\mathsf{fi}}$, we have

$$W_{\mathsf{L}}(m) = \frac{\lambda_{\mathsf{L}}}{Z_m}\sum_{k=0}^{m-1}\lambda^k u_{\mathsf{L}}(k),$$

where $Z_m = \sum_{k=0}^{m-1}\lambda^k + \sum_{k=m}^{\infty}\lambda^m\lambda_{\mathsf{H}}^{k-m} = \frac{1-\lambda^m}{1-\lambda} + \frac{\lambda^m}{1-\lambda_{\mathsf{H}}}$. Thus, we obtain

$$\begin{aligned}
W_{\mathsf{L}}(m-1) - W_{\mathsf{L}}(m) &= \frac{\lambda_{\mathsf{L}}}{Z_{m-1}}\sum_{k=0}^{m-2}\lambda^k u_{\mathsf{L}}(k) - \frac{\lambda_{\mathsf{L}}}{Z_m}\sum_{k=0}^{m-1}\lambda^k u_{\mathsf{L}}(k)\\
&= \frac{\lambda_{\mathsf{L}}}{Z_{m-1}}\sum_{k=0}^{m-2}\lambda^k u_{\mathsf{L}}(k) - \frac{\lambda_{\mathsf{L}}}{Z_{m-1}}\sum_{k=0}^{m-1}\lambda^k u_{\mathsf{L}}(k) + \frac{\lambda_{\mathsf{L}}}{Z_{m-1}}\sum_{k=0}^{m-1}\lambda^k u_{\mathsf{L}}(k) - \frac{\lambda_{\mathsf{L}}}{Z_m}\sum_{k=0}^{m-1}\lambda^k u_{\mathsf{L}}(k)\\
&= \frac{\lambda_{\mathsf{L}}\lambda^{m-1}}{(1-\lambda_{\mathsf{H}})Z_{m-1}}\left(W_{\mathsf{L}}(m) - (1-\lambda_{\mathsf{H}})u_{\mathsf{L}}(m-1)\right)\\
&= \left(\frac{(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m-1}}\right)\left(W_{\mathsf{L}}(m) - (1-\lambda_{\mathsf{H}})u_{\mathsf{L}}(m-1)\right).
\end{aligned}$$

Equivalently, we have

$$W_{\mathsf{L}}(m-1) = \left(1 + \frac{(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m-1}}\right)W_{\mathsf{L}}(m) - \left(\frac{(1-\lambda_{\mathsf{H}})(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m-1}}\right)u_{\mathsf{L}}(m-1).$$

Letting $m = m_{\mathsf{fi}}$ and using the assumption that $u_{\mathsf{L}}(m_{\mathsf{fi}}-1) \leq W_{\mathsf{L}}(\mathsf{fi})$, we get

$$W_{\mathsf{L}}(m_{\mathsf{fi}}-1) \geq \left(1 + \frac{\lambda_{\mathsf{H}}(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m_{\mathsf{fi}}-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m_{\mathsf{fi}}-1}}\right)W_{\mathsf{L}}(m_{\mathsf{fi}}) = \beta_{\mathsf{L},\mathsf{sm}}\cdot W_{\mathsf{L}}(m_{\mathsf{fi}}).$$

Note that by definition, $1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m_{\mathsf{fi}}-1} > 1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}} > 0$ for $\lambda < 1$, implying that $\beta_{\mathsf{L},\mathsf{sm}} > 1$.

Next, we proceed to the H type. Similar to $W_{\mathsf{L}}(m)$, we have

$$W_{\mathsf{H}}(m) = \frac{\lambda_{\mathsf{H}}}{Z_m}\left(\sum_{k=0}^{m-1}\lambda^k u_{\mathsf{H}}(k) + \lambda^m\sum_{k=m}^{\infty}\lambda_{\mathsf{H}}^{k-m}u_{\mathsf{H}}(k)\right) \triangleq \lambda_{\mathsf{H}}\frac{F_m}{Z_m}.$$

Thus, we get

$$W_{\mathsf{H}}(m-1) - W_{\mathsf{H}}(m) = \frac{\lambda_{\mathsf{H}}}{Z_{m-1}}(F_{m-1} - F_m) + \frac{\lambda_{\mathsf{L}}\lambda^{m-1}}{(1-\lambda_{\mathsf{H}})Z_{m-1}}W_{\mathsf{H}}(m).$$

Furthermore,

$$F_{m-1} - F_m = -\lambda_{\mathsf{L}}\lambda^{m-1}\sum_{k=m}^{\infty}\lambda_{\mathsf{H}}^{k-m}u_{\mathsf{H}}(k).$$

Thus,

$$W_{\mathsf{H}}(m-1)-W_{\mathsf{H}}(m)=\frac{\lambda_{\mathsf{L}}\lambda^{m-1}}{(1-\lambda_{\mathsf{H}})Z_{m-1}}\left(W_{\mathsf{H}}(m)-\lambda_{\mathsf{H}}\left(\sum_{k=m}^{\infty}(1-\lambda_{\mathsf{H}})\lambda_{\mathsf{H}}^{k-m}u_{\mathsf{H}}(k)\right)\right)$$
$$=\left(\frac{(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m-1}}\right)\left(W_{\mathsf{H}}(m)-\lambda_{\mathsf{H}}\left(\sum_{k=m}^{\infty}(1-\lambda_{\mathsf{H}})\lambda_{\mathsf{H}}^{k-m}u_{\mathsf{H}}(k)\right)\right).$$

This implies that

$$W_{\mathsf{H}}(m-1)=\left(1+\frac{(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m-1}}\right)W_{\mathsf{H}}(m)-\left(\frac{\lambda_{\mathsf{H}}(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m-1}}\right)\left(\sum_{k=m}^{\infty}(1-\lambda_{\mathsf{H}})\lambda_{\mathsf{H}}^{k-m}u_{\mathsf{H}}(k)\right)$$
$$\geq\left(1+\frac{(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m-1}}\right)W_{\mathsf{H}}(m)-\left(\frac{\lambda_{\mathsf{H}}(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m-1}}\right)u_{\mathsf{H}}(m).$$

Again, letting $m=m_{\mathsf{fi}}$ and using the assumption that $u_{\mathsf{H}}(m_{\mathsf{fi}})\leq W_{\mathsf{H}}(\mathsf{fi})$, we get

$$W_{\mathsf{H}}(m_{\mathsf{fi}}-1)\geq\left(1+\frac{(1-\lambda_{\mathsf{H}})(1-\lambda)\lambda_{\mathsf{L}}\lambda^{m_{\mathsf{fi}}-1}}{1-\lambda_{\mathsf{H}}-\lambda_{\mathsf{L}}\lambda^{m_{\mathsf{fi}}-1}}\right)W_{\mathsf{H}}(m_{\mathsf{fi}})=\beta_{\mathsf{H},\mathsf{sm}}\cdot W_{\mathsf{H}}(m_{\mathsf{fi}}).$$

Before proving the third part of the proposition, we note that using Assumption 1, along with a stochastic dominance argument, we obtain that function $g(x)\triangleq\sum_{k\in\mathbb{N}_0}(1-x)x^k u_{\mathsf{L}}(k)$ is strictly decreasing in $x\in[0,1)$. Furthermore, $g(0)=u_{\mathsf{L}}(0)>0$ and $\lim_{x\to1^-}g(x)<0$. Thus, there exists a unique $\bar{\Lambda}_{\mathsf{H}}\in(0,1)$ such that $g(\bar{\Lambda}_{\mathsf{H}})=0$. To prove the second part of the proposition, we show that if $\lambda_{\mathsf{H}}\geq\bar{\Lambda}_{\mathsf{H}}$, then no L-type user joins under the no-information mechanism, i.e., $p^{\mathsf{ni}}=0$. The result then follows from Theorem 5. Thus, suppose $\lambda_{\mathsf{H}}\geq\bar{\Lambda}_{\mathsf{H}}$, and no (other) L-type user joins the queue under no-information mechanism. The steady-state distribution π of the queue is then that of an $M/M/1$ queue with arrival rate λ_{H}, and hence we have $\pi_n=(1-\lambda_{\mathsf{H}})\lambda_{\mathsf{H}}^n$ for $n\geq0$. This implies that the expected utility (in steady-state) of a L-type user for joining is given by $\sum_{k\in\mathbb{N}_0}\pi_n u_{\mathsf{L}}(n)=\sum_{k\in\mathbb{N}_0}(1-\lambda_{\mathsf{H}})\lambda_{\mathsf{H}}^k u_{\mathsf{L}}(n)=g(\lambda_{\mathsf{H}})\leq0$. The inequality follows from the fact that $g(x)$ is decreasing in x and equals zero when $x=\bar{\Lambda}_{\mathsf{H}}$. Thus, we obtain that the optimal action for a L-type user is indeed not to join, and hence $p^{\mathsf{ni}}=0$. This completes the proof. □

Proof of Theorem 4. Recall that under the full-information mechanism fi, the L-type users receive the "join" signal if and only if the queue-length is strictly less than the full-information threshold m_{fi}. Thus, conditional on receiving the "leave" signal, the queue-length is at least m_{fi}, and the expected utility of the L-type users for joining the queue is given by

$$U_{\mathsf{L}}(0,\mathsf{join})\leq u_{\mathsf{L}}(m_{\mathsf{fi}})<0,$$

where we have used the definition of m_{fi} and the fact that $u_{\mathsf{L}}(k)$ is strictly decreasing in k. Together with the fact that the probability of receiving a "leave" signal is positive under the full-information mechanism, we obtain that $L(m_{\mathsf{fi}})<0$, and hence the (LEAVE) condition does not bind. Hence,

from Theorem 3, we conclude that the full-information mechanism is Pareto-efficient within the class Π_{AP} if and only if it is so within the class Π_{SM}. Thus, we obtain the dichotomy in the theorem statement.

To show the final part of the theorem, suppose the full-information mechanism is Pareto-efficient within the class of admission policies Π_{AP}. Consider the admission policy with threshold $m_{\mathsf{fi}}-1$. In Lemma EC.3 (see Appendix EC.2), we show that $W_{\mathsf{H}}(x)$ is strictly decreasing in the threshold x. Hence, we have $W_{\mathsf{H}}(m_{\mathsf{fi}}-1) > W_{\mathsf{H}}(m_{\mathsf{fi}})$. Since the full-information mechanism is Pareto-efficient within Π_{AP}, this implies $W_{\mathsf{L}}(m_{\mathsf{fi}}) > W_{\mathsf{L}}(m_{\mathsf{fi}}-1)$.

Conversely, suppose $W_{\mathsf{L}}(m_{\mathsf{fi}}) > W_{\mathsf{L}}(m_{\mathsf{fi}}-1)$. In Lemma EC.3, we also show that $W_{\mathsf{L}}(x)$ is unimodal, i.e, $W_{\mathsf{L}}(x)$ is increasing for small $x \in \mathbb{R}_+$ and decreasing otherwise. The unimodality then implies that for all $0 \le \hat{x} < m_{\mathsf{fi}}$, we have $W_{\mathsf{L}}(\hat{x}) < W_{\mathsf{L}}(m_{\mathsf{fi}})$. Thus, no admission policy with threshold $\hat{x} < m_{\mathsf{fi}}$ Pareto dominates the full-information mechanism. Since any admission policy that is not Pareto-efficient is dominated by some threshold policy with threshold less than or equal to m_{fi}, we obtain that the full-information mechanism, with threshold m_{fi}, is Pareto-efficient within the class Π_{AP} of admission policies. □

Proof of Theorem 5. Recall that under the no-information mechanism, a L-type users joins the queue with a fixed probability $p^{\mathsf{ni}} \in [0,1]$ irrespective of the queue-length upon arrival.

For $p^{\mathsf{ni}} \in (0,1)$ it is straightforward to verify that the resulting steady-state distribution does not have a threshold structure, and hence by Theorem 1, the no-information mechanism is not Pareto-efficient. For $p^{\mathsf{ni}} = 1$, the resulting steady-state distribution has a threshold structure with threshold equal to infinity. In this case, Theorem 1 implies that the no-information mechanism is not Pareto-efficient. Thus, if $p^{\mathsf{ni}} \in (0,1]$, then the no-information mechanism is Pareto-dominated within the class Π_{SM} of signaling mechanisms.

Finally, suppose $p^{\mathsf{ni}} = 0$. Then, the steady-state distribution π^{ni} is given by $\pi^{\mathsf{ni}}_n = (1-\lambda_{\mathsf{H}})\lambda_{\mathsf{H}}^n$ for $n \ge 0$. Now, consider any other admission policy $\hat{\pi} \in \Pi_{\mathsf{AP}}$, where at least some fraction of L-type users are admitted into the queue. Using a coupling argument, it is straightforward to show that $\hat{\pi}$ stochastically dominates π^{ni}. Since $u_{\mathsf{H}}(n)$ is strictly decreasing in n, this further implies that $W_{\mathsf{H}}(\hat{\pi}) < W_{\mathsf{H}}(\pi^{\mathsf{ni}})$. Hence, it follows that the no-information mechanism ni is Pareto-efficient within the class Π_{AP} of admission policies. □

EC.4. Proofs from Section 5

Proof of Theorem 6. First, suppose $\lambda_{\mathsf{H}} \in [\bar{\Lambda}_{\mathsf{H}}, 1]$, and fix a $\theta \in [0,1]$. From Theorem 5, we obtain that the no-information mechanism ni is Pareto-efficient, and furthermore, under ni, all L-type users choose their outside option. Consider the admission policy $\mathsf{ap}(\theta)$. If $\mathsf{ap}(\theta)$ makes some L-type users join the queue, then the welfare of H-type users can only be lower than that in ni:

$W_{\mathsf{H}}(\mathsf{ni}) \geq W_{\mathsf{H}}(\pi)$. Thus, for $\mathsf{ap}(\theta)$ to be Pareto-efficient, we must have $W_{\mathsf{L}}(\mathsf{ap}(\theta)) > W_{\mathsf{L}}(\mathsf{ni}) = 0$. Thus, we have $J(\mathsf{ap}(\theta)) = W_{\mathsf{L}}(\mathsf{ap}(\theta)) > 0$, and hence the obedience constraint (JOIN) holds. Furthermore, we have

$$J(\mathsf{ap}(\theta)) + L(\mathsf{ap}(\theta)) = \lambda_{\mathsf{L}} \sum_{n \in \mathbb{N}_0} \pi_n(\mathsf{ap}(\theta)) u_{\mathsf{L}}(n) \leq 0,$$

where $\pi(\mathsf{ap}(\theta))$ denotes the steady-state distribution under $\mathsf{ap}(\theta)$. This is because $\pi(\mathsf{ap}(\theta))$ stochastically dominates the steady-state distribution under ni, and hence the right-hand side expression is less than $\lambda_{\mathsf{L}} \sum_{n \in \mathbb{N}_0} (1 - \lambda_{\mathsf{H}}) \lambda_{\mathsf{H}}^n u_{\mathsf{L}}(n)$, which is non-positive as $\lambda_{\mathsf{H}} \geq \bar{\Lambda}_{\mathsf{H}}$. Since $J(\mathsf{ap}(\theta)) \geq 0$, this implies that $L(\mathsf{ap}(\theta)) \leq 0$, and hence $\mathsf{ap}(\theta)$ also satisfies the obedience constraint LEAVE. Taken together, we obtain that $\mathsf{ap}(\theta) \in \Pi_{\mathsf{SM}}$, and hence $\mathsf{ap}(\theta) = \mathsf{sm}(\theta)$.

Next, let $\lambda_{\mathsf{H}} < \bar{\Lambda}_{\mathsf{H}}$. Fix $\theta_1, \theta_2 \in [0,1]$ with $\theta_2 > \theta_1$, and let x_i denote the threshold of the Pareto-efficient admission policy $\mathsf{ap}(\theta_i)$. In the following, we first show that $x_1 \leq x_2$. By the definition of $W(\pi, \theta)$ and $\mathsf{ap}(\theta)$, we have

$$\theta_1 W_{\mathsf{L}}(x_1) + (1 - \theta_1) W_{\mathsf{H}}(x_1) \geq \theta_1 W_{\mathsf{L}}(x_2) + (1 - \theta_1) W_{\mathsf{H}}(x_2),$$
$$\theta_2 W_{\mathsf{L}}(x_2) + (1 - \theta_2) W_{\mathsf{H}}(x_2) \geq \theta_2 W_{\mathsf{L}}(x_1) + (1 - \theta_2) W_{\mathsf{H}}(x_1).$$

After some algebra, we obtain

$$W_{\mathsf{L}}(x_2) - W_{\mathsf{L}}(x_1) \geq W_{\mathsf{H}}(x_2) - W_{\mathsf{H}}(x_1).$$

Now, if $x_1 > x_2$, then from Lemma EC.3 in Appendix EC.2, we obtain $W_{\mathsf{H}}(x_1) < W_{\mathsf{H}}(x_2)$. The preceding inequality would then imply $W_{\mathsf{L}}(x_1) < W_{\mathsf{L}}(x_2)$. However, this would imply that the admission policy $\mathsf{ap}(\theta_1)$ is Pareto-dominated by the policy $\mathsf{ap}(\theta_2)$, a contradiction. Thus, we obtain that $x_1 \leq x_2$.

Next, suppose the admission policy $\mathsf{ap}(\theta_1)$ satisfies the obedience constraints, and hence $L(x_1) \leq 0$. In Lemma EC.4 (stated and proven in Appendix EC.2), we establish that if $L(x) \leq 0$ then $L(u) \leq 0$ for all $u \geq x$. Since $x_1 \leq x_2$, Lemma EC.4 implies that $L(x_2) \leq 0$, and hence the (LEAVE) condition holds for $\mathsf{ap}(\theta_2)$. Further, by Theorem 2 we have $x_2 \leq m_{\mathsf{fi}}$, which implies $J(x_2) \geq 0$ and hence the (JOIN) condition holds for $\mathsf{ap}(\theta_2)$. Together, we obtain that $\mathsf{ap}(\theta_2)$ also satisfies the obedience constraints.

Thus, we conclude that if for some $\theta_1 \in [0,1]$ the admission policy $\mathsf{ap}(\theta_1)$ satisfies the obedience constraints, then so does the admission policy $\mathsf{ap}(\theta_2)$ for all $\theta_2 > \theta_1$. This implies the existence of (a smallest such) $\theta(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}}) \in [0,1]$ such that for all $\theta > \theta(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}})$ we have $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$.[18]

[18] Note that in this case, the threshold of the admission policy $\mathsf{ap}(\theta)$ (or equivalently the signaling mechanism $\mathsf{sm}(\theta)$) can be positive. For numerical examples, see Figure 4 and its related discussion in Section 6.

(Note that we allow the possibility that $\theta(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}}) = 1$.) Further, we have $\theta(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}}) > 0$, since for $\theta = 0$, the admission policy $\mathsf{ap}(0)$ makes all L-type users take the outside option. However, the obedience condition (LEAVE) does not hold for $\mathsf{ap}(0)$ since $\lambda_{\mathsf{H}} < \bar{\Lambda}_{\mathsf{H}}$.

Finally, for $\theta < \theta(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}})$, the admission policy $\mathsf{ap}(\theta)$ does not satisfy the obedience constraints, and hence $\mathsf{sm}(\theta) \neq \mathsf{ap}(\theta)$. Theorem 3 then implies that the (LEAVE) condition binds for all such θ, i.e., $L(\mathsf{sm}(\theta)) = 0$. In Lemma EC.4, we also prove that $L(x)$ is strictly decreasing as long as it is non-negative, and remains negative subsequently. Thus, there exists a unique threshold $\bar{x} \leq m_{\mathsf{fi}}$ (independent of θ) with $L(\bar{x}) = 0$. From this, we conclude that $\mathsf{sm}(\theta)$ is the threshold mechanism with threshold $\bar{x}$ for all $\theta < \theta(\lambda_{\mathsf{L}}, \lambda_{\mathsf{H}})$. $\square$

EC.5. Proofs from Section 7.1

Proof of Proposition 4. Proof of Part 1: Let π denote the full-information mechanism. Suppose π is Pareto dominated by an admission policy $\tilde{\pi}$, i.e., $W_{\mathsf{H}}(\tilde{\pi}) \geq W_{\mathsf{H}}(\pi)$ and $W_{\mathsf{L}}(\tilde{\pi}) \geq W_{\mathsf{L}}(\pi)$, with at least one inequality strict. Since under the full-information mechanism π, none of the obedience constraints bind, we obtain that for small enough $\delta > 0$, the admission policy $\pi_\delta = (1-\delta)\pi + \delta\tilde{\pi}$ satisfies all the obedience constraints, and hence can be implemented as a signaling mechanism. Using the linearity of the welfare functions, we conclude that π_δ Pareto dominates the full-information mechanism π.

Finally, using Lemma EC.5 stated later in this section, we obtain $W_{\mathsf{L}}(m_{\mathsf{L}}, m_{\mathsf{H}} - 1) > W_{\mathsf{L}}(m_{\mathsf{L}}, m_{\mathsf{H}})$ and $W_{\mathsf{H}}(m_{\mathsf{L}} - 1, m_{\mathsf{H}}) \geq W_{\mathsf{H}}(m_{\mathsf{L}}, m_{\mathsf{H}})$. Thus, if $W_{\mathsf{H}}(m_{\mathsf{L}}, m_{\mathsf{H}} - 1) \geq W_{\mathsf{H}}(m_{\mathsf{L}}, m_{\mathsf{H}})$, then we obtain that the threshold mechanism $\mathsf{Th}(m_{\mathsf{L}}, m_{\mathsf{H}} - 1)$ Pareto-dominates the full information mechanism. On the other hand, if $W_{\mathsf{L}}(m_{\mathsf{L}} - 1, m_{\mathsf{H}}) \geq W_{\mathsf{L}}(m_{\mathsf{L}}, m_{\mathsf{H}})$, then the threshold mechanism $\mathsf{Th}(m_{\mathsf{L}} - 1, m_{\mathsf{H}})$ Pareto dominates the full information. In either case, there exists a threshold signaling mechanism that Pareto dominates the full-information mechanism, since for any $\delta > 0$, the signaling mechanisms $(1-\delta)\mathsf{Th}(m, n-1) + \delta\mathsf{Th}(m, n)$ and $(1-\delta)\mathsf{Th}(m-1, n) + \delta\mathsf{Th}(m, n)$ both have a threshold structure.

Proof of Part 2: We begin by showing that the no-information mechanism is Pareto dominated in the class of admission policies. First, suppose under the no-information mechanism, the L-type users never join, while the H-type users join with some probability $p \in (0, 1]$. In this case, the admission policy that never admits the L-type, and implements the admission rule that maximizes the H types' welfare Pareto dominates the no-information mechanism. Next, if the L-type users join with probability $p \in (0, 1)$ under the no-information mechanism, then due to the assumption on the utilities, the H-type user always joins. Since $p \in (0, 1)$, the welfare of the L-type in this case is zero. This implies that the admission policy that never admits L-type user and always admits the H-type user Pareto dominates the no-information mechanism. Finally, suppose both types join with probability 1 under the no-information mechanism. In this case, the admission policy that

never admits any type above queue length m_{H} and always admits below this queue length achieves higher utility for both types, and hence Pareto dominates the no-information mechanism.

Next, suppose under the no-information mechanism, the L-type users join with positive probability. To show that the no-information mechanism is Pareto dominated by a signaling mechanism, we split the argument into two cases:

1. Suppose in equilibrium, both types join with probability 1. Consider the threshold mechanism $\mathsf{Th}(m_{\mathsf{H}}, m_{\mathsf{H}})$, i.e., the mechanism sends signal 2 up to queue length m_{H}, and sends signal 0 afterwards. From a straightforward argument, it follows that this mechanism is obedient, and achieves higher welfare for both types than the no-information mechanism.
2. Suppose in equilibrium, the H-type users join with probability 1 and the L-type users join with probability $p \in (0,1)$. Letting π denote the no-information mechanism, we have $\pi_{k,0} = 0$, $\pi_{k,1} > 0$ and $\pi_{k,2} > 0$ for all $k \geq 0$. Furthermore, we have $S_{\mathsf{H},1}(\pi) > S_{\mathsf{L},1}(\pi) = 0$. Thus, by Lemma EC.6 stated later in this section, we obtain that no-information mechanism is Pareto dominated by a signaling mechanism.

Taken together, we obtain the result. □

The following lemmas are used in the proof of Proposition 4.

LEMMA EC.5. *Suppose $m \leq m_{\mathsf{L}}$. Then, for $n \geq m$ we have $W_{\mathsf{L}}(m, n-1) > W_{\mathsf{L}}(m,n)$ and $W_{\mathsf{H}}(m-1,n) > W_{\mathsf{H}}(m,n)$.*

Proof. First we define two auxiliary functions: $\Psi(m) \triangleq W_{\mathsf{L}}(m,n)/Z(m,n)$ and $\Phi(m,n) \triangleq W_{\mathsf{H}}(m,n)/Z(m,n)$ where $Z(m,n)$, $W_{\mathsf{L}}(m,n)$, $W_{\mathsf{H}}(m,n)$ are as defined in footnotes 15 and 16.

Let $m \leq m_{\mathsf{L}}$ and $n \geq m \in \mathbb{N}_0$. We have

$$\begin{aligned} W_{\mathsf{L}}(m,n-1) - W_{\mathsf{L}}(m,n) &= (Z(m,n-1) - Z(m,n))\Psi(m) \\ &= Z(m,n)Z(m,n-1)\left(\frac{1}{Z(m,n)} - \frac{1}{Z(m,n-1)}\right)\Psi(m) \\ &= Z(m,n-1)W_{\mathsf{L}}(m,n)\lambda^m \lambda_{\mathsf{H}}^{n-m}. \end{aligned}$$

Since $m \leq m_{\mathsf{L}}$, we have $W_{\mathsf{L}}(m,n) > 0$. Thus, we obtain $W_{\mathsf{L}}(m,n-1) > W_{\mathsf{L}}(m,n)$. Next, we have

$$\begin{aligned} &W_{\mathsf{H}}(m-1,n) - W_{\mathsf{H}}(m,n) \\ &= Z(m-1,n)\Phi(m-1,n) - Z(m,n)\Phi(m,n) \\ &= Z(m-1,n)\left(\Phi(m-1,n) - \Phi(m,n)\right) + \Phi(m,n)\left(Z(m-1,n) - Z(m,n)\right) \\ &= Z(m-1,n)Z(m,n)\lambda_{\mathsf{L}}\lambda^{m-1}\lambda_{\mathsf{H}} \cdot \\ &\quad \left(\left(\sum_{k=0}^{m-1}\lambda^k u_{\mathsf{H}}(k)\right)\left(\sum_{k=m}^{n}\lambda_{\mathsf{H}}^{k-m}\right) - \left(\sum_{k=0}^{m-1}\lambda^k\right)\left(\sum_{k=m}^{n-1}\lambda_{\mathsf{H}}^{k-m}u_{\mathsf{H}}(k)\right)\right) \end{aligned}$$

$$\geq Z(m-1,n)Z(m,n)\lambda_{\mathsf{L}}\lambda^{m-1}\lambda_{\mathsf{H}}\left(\sum_{k=0}^{m-1}\lambda^k\right)\left(\lambda_{\mathsf{H}}^{n-m}u_{\mathsf{H}}(m-1)+\left(\sum_{k=m}^{n-1}\lambda_{\mathsf{H}}^{k-m}\right)(u_{\mathsf{H}}(m-1)-u_{\mathsf{H}}(m))\right)$$
$$>0,$$

where the final inequality follows from the fact that u_{H} is strictly decreasing, and since $m \leq m_{\mathsf{L}}$, we have $u_{\mathsf{H}}(m-1) \geq u_{\mathsf{L}}(m-1) > 0$. $\square$

LEMMA EC.6. *Consider a signaling mechanism $\pi = \{\pi_{k,j} : j = 0,1,2; k \geq 0\}$ such that $\pi_{k,0} = 0$ for all $k \geq 0$ and there exists an $m \geq 0$ with $\pi_{m,1} > 0$ and $\pi_{m+1,2} > 0$. If in addition $S_{\mathsf{H},1}(\pi) > 0$, then π is Pareto dominated by a signaling mechanism.*

Proof. Suppose π is as stated in the lemma statement, and furthermore, there exists an $m \geq 0$ such that $\pi_{m,1} > 0$ and $\pi_{m+1,2} > 0$. Then, for small enough $\delta > 0$, define $\tilde{\pi}$ as follows:

$$\tilde{\pi}_{k,2} = \begin{cases} \pi_{k,2} & \text{for } k < m; \\ \pi_{m,2} + \frac{\delta}{\lambda_{\mathsf{L}}}\sum_{n>m+1}\sum_j \pi_{n,j} & \text{for } k = m; \\ (1-\delta)\pi_{m+1,2} - \frac{\delta\lambda_{\mathsf{H}}}{\lambda_{\mathsf{L}}}\sum_{n\geq m+1}\sum_j \pi_{n,j} & \text{for } k = m+1; \\ (1-\delta)\pi_{k,2} & \text{for } k > m+1; \end{cases}$$
$$\tilde{\pi}_{k,1} = \begin{cases} \pi_{k,1} & \text{for } k < m; \\ \pi_{m,1} - \frac{\delta}{\lambda_{\mathsf{L}}}\sum_{n>m+1}\sum_j \pi_{n,j} & \text{for } k = m; \\ (1-\delta)\pi_{m+1,1} + \frac{\delta\lambda}{\lambda_{\mathsf{L}}}\sum_{n\geq m+1}\sum_j \pi_{n,j} & \text{for } k = m+1; \\ (1-\delta)\pi_{k,1} & \text{for } k > m+1; \end{cases}$$
$$\tilde{\pi}_{k,0} = \pi_{k,0} - \delta\pi_{k,0}\mathbf{I}\{k \geq m+1\}.$$

Then, it is straightforward to verify that $\tilde{\pi}$ satisfies the balance conditions, given by $\lambda\pi_{k,2} + \lambda_{\mathsf{H}}\pi_{k,1} = \sum_j \pi_{k+1,j}$ for all $k \geq 0$. Furthermore, we have, for each $i \in \{\mathsf{H},\mathsf{L}\}$,

$$\begin{aligned} S_{i,2}(\tilde{\pi}) - S_{i,2}(\pi) &= \frac{\delta}{\lambda_{\mathsf{L}}}\sum_{n>m+1}\left(\textstyle\sum_j \pi_{n,j}\right)(u_i(m) - u_i(n-1) - \lambda_{\mathsf{H}}(u_i(m+1) - u_i(n))) \\ &\quad - \frac{\delta\lambda_{\mathsf{H}}}{\lambda_{\mathsf{L}}}\sum_{n\geq m+1}\pi_{n,0}u_i(n) \\ S_{i,1}(\tilde{\pi}) - S_{i,1}(\pi) &= -\frac{\delta}{\lambda_{\mathsf{L}}}\sum_{n>m+1}\left(\textstyle\sum_j \pi_{n,j}\right)(u_i(m) - u_i(n-1) - \lambda(u_i(m+1) - u_i(n))) \\ &\quad + \frac{\delta\lambda}{\lambda_{\mathsf{L}}}\sum_{n\geq m+1}\pi_{n,0}u_i(n) \\ S_{i,0}(\tilde{\pi}) - S_{i,0}(\pi) &= -\delta\sum_{n\geq m+1}\pi_{n,0}u_i(n). \end{aligned}$$

Now, for any $a \in [0,1)$, we have for all $n > m+1$,

$$\begin{aligned} u_i(m) - u_i(n-1) - a(u_i(m+1) - u_i(n)) &> u_i(m) - u_i(n-1) - (u_i(m+1) - u_i(n)) \\ &= u_i(m) - u_i(m+1) - (u_i(n-1) - u_i(n)) \geq 0, \end{aligned}$$

where the first inequality follows from the fact that $u_i(n)$ is strictly decreasing, and the second inequality follows from the fact that $u_i(k)-u_i(k+1)$ is non-increasing. Furthermore, since $\pi_{m+1,2}>0$, we must have $\sum_j \pi_{m+2,j}>0$. Coupled with the fact that $\pi_{k,0}=0$ for all $k\geq 0$, we obtain for all small enough $\delta>0$ and for $i\in\{\mathsf{H},\mathsf{L}\}$,

$$S_{\mathsf{L},2}(\tilde{\pi})>S_{\mathsf{L},2}(\pi), \qquad S_{\mathsf{L},1}(\tilde{\pi})<S_{\mathsf{L},1}(\pi), \qquad S_{i,0}(\tilde{\pi})=S_{i,0}(\pi).$$

Since π is obedient with $S_{H,1}(\pi)>0$, we conclude that $\tilde{\pi}$ is obedient as well for small enough $\delta>0$. Finally, we have

$$\begin{aligned}
W_{\mathsf{L}}(\tilde{\pi}) &= \lambda_{\mathsf{L}} S_{\mathsf{L},2}(\tilde{\pi}) > \lambda_{\mathsf{L}} S_{\mathsf{L},2}(\pi) = W_{\mathsf{L}}(\pi)\\
W_{\mathsf{H}}(\tilde{\pi}) &= \lambda_{\mathsf{H}}(S_{\mathsf{H},1}(\tilde{\pi})+S_{\mathsf{H},2}(\tilde{\pi}))\\
&= \lambda_{\mathsf{H}}\left(S_{\mathsf{H},1}(\pi)+S_{\mathsf{H},2}(\pi)\right)\\
&\quad + \delta \sum_{n>m+1}\left(\textstyle\sum_j \pi_{n,j}\right)\left(u_{\mathsf{H}}(m+1)-u_{\mathsf{H}}(n)\right) + \delta \sum_{n\geq m+1}\pi_{n,0}u_{\mathsf{H}}(n)\\
&> W_{\mathsf{H}}(\pi),
\end{aligned}$$

where the final inequality follows from the fact that u_{H} is strictly decreasing and $\sum_j \pi_{m+2,j}>0$. Thus, we obtain that π is Pareto dominated by $\tilde{\pi}$.

Thus, we obtain that any π with $\pi_{k,0}=0$, $S_{\mathsf{H},1}(\pi)>0$ and for which there exists an $m\geq 0$ such that $\pi_{m,1}>0$ and $\pi_{m+1,2}>0$ cannot be Pareto efficient. $\square$

EC.6. Further Numerical Analysis for Section 7.1

In this section, we numerically examine the structure of the optimal signaling mechanism in the fully persuadable population setting introduced in Section 7.1. We start by presenting two examples which show that $\mathsf{sm}(\theta)$ may not have the structure of a threshold mechanism as defined in Section 7.1.

Examples: Suppose $u_{\mathsf{L}}(k)=1-c(k+1)$ and $u_{\mathsf{H}}(k)=1-c(k+1)-\ell_{\mathsf{H}}$ with $c=0.15$, and $\ell_{\mathsf{H}}=-0.7$. Further, let $\lambda_{\mathsf{H}}=0.7$ and $\lambda_{\mathsf{L}}=1-\lambda_{\mathsf{H}}=0.3$. Recall that $\sigma(n,s)\in[0,1]$ denotes the probability of sending signal $s\in\{0,1,2\}$ when the queue length is n. By solving the linear program introduced in Section 7.1, we obtain that:

1. The mechanism $\mathsf{sm}(0.7)$ is given by:

$$\sigma(n,s)=\begin{cases}
\mathbf{I}(s=2) & \text{for } 0\leq n<4;\\
\mathbf{I}(s=1) & \text{for } 4\leq n<10;\\
0.444\times\mathbf{I}(s=1)+0.556\times\mathbf{I}(s=0) & \text{for } n=10;\\
0.190\times\mathbf{I}(s=1)+0.810\times\mathbf{I}(s=0) & \text{for } n=11;\\
\mathbf{I}(s=0) & \text{otherwise,}
\end{cases}$$

2. The mechanism $\mathsf{sm}(0.8)$ is given by:

$$\sigma(n,s)=\begin{cases}\mathbf{I}(s=2) & \text{for } 0\le n<4;\\ \mathbf{I}(s=1) & \text{for } 4\le n<9;\\ 0.774\times\mathbf{I}(s=1)+0.226\times\mathbf{I}(s=0) & \text{for } n=9;\\ \mathbf{I}(s=1) & \text{for } n=10;\\ 0.199\times\mathbf{I}(s=1)+0.801\times\mathbf{I}(s=0) & \text{for } n=11;\\ \mathbf{I}(s=0) & \text{otherwise,}\end{cases}$$

The above examples show that while the signaling mechanism still follows a "monotone" structure by sending signal 2 (i.e., join for both types) for small queue length, and then signal 1 (i.e., leave for L-type and join for H-type) for medium queue length and then signal 0 (i.e., leave for both types) for sufficiently large queue lengths, the queue length at which the mechanism randomizes between the two signals does not necessarily follow the structure of the $\mathsf{Th}(x,y)$ defined at the beginning of Section 7.1.

Even though the above examples show that the optimal signaling can be "slightly" different from $\mathsf{Th}(x,y)$, our numerical analysis confirms that there will be little loss in limiting ourselves to the class of threshold signaling mechanism. As a representative example, for model primitives: $\lambda_{\mathsf{L}}=1-\lambda_{\mathsf{H}}$ with $\lambda_{\mathsf{H}}\in[0,1]$, $\ell_{\mathsf{H}}=-0.7$, $u_{\mathsf{L}}(k)=1-c(k+1)$, and $u_{\mathsf{H}}(k)=1-c(k+1)-\ell_{\mathsf{H}}$ with $c=0.15$, we compute, for each $(\theta,\lambda_{\mathsf{H}})$, the best threshold signaling mechanism (found through exhaustive search on a grid of two thresholds with 1/16 increments) which we denote by tsm. In Figure EC.1, we plot the heat map of $\frac{W(\mathsf{sm},\theta)-W(\mathsf{tsm},\theta)}{W(\mathsf{sm},\theta)-W(\mathsf{fi},\theta)}$. (Note that we use $W(\mathsf{sm},\theta)-W(\mathsf{fi},\theta)$ as the normalization factor to ensure that the ratio is in $[0,1]$.) We observe that the normalized gap is zero for most values of $(\theta,\lambda_{\mathsf{H}})$; in the regime of $(\theta,\lambda_{\mathsf{H}})$ where the gap is nonzero—which includes the examples presented above—it is very small and notably far from 1. Thus our numerical analysis suggests that if for practical reasons, using a threshold mechanism is more desirable, there exists a threshold mechanism which performs nearly as well as the optimal one, and better than the full-information mechanism.

Our further numerical analysis—which we omit for the sake of brevity—shows that in the linear utility case, such deviations from a threshold mechanism only occurs when $|\ell_{\mathsf{H}}|$ is small. For example, for model primitives used in Figure 5 where $\ell_{\mathsf{H}}\in\{-1,-5,-10\}$, for any $\theta\in\{1/12,2/12,\dots,11/12\}$ the optimal signaling mechanism has a threshold structure.

EC.7. Further Numerical Analysis for Section 7.2

In this section, we expand our numerical analysis for the model introduced in Section 7.2, where the two types have different service rates. First, using the linear program developed in Section 7.2 we verify the power of information design for a wide range of gap between the two service rates.

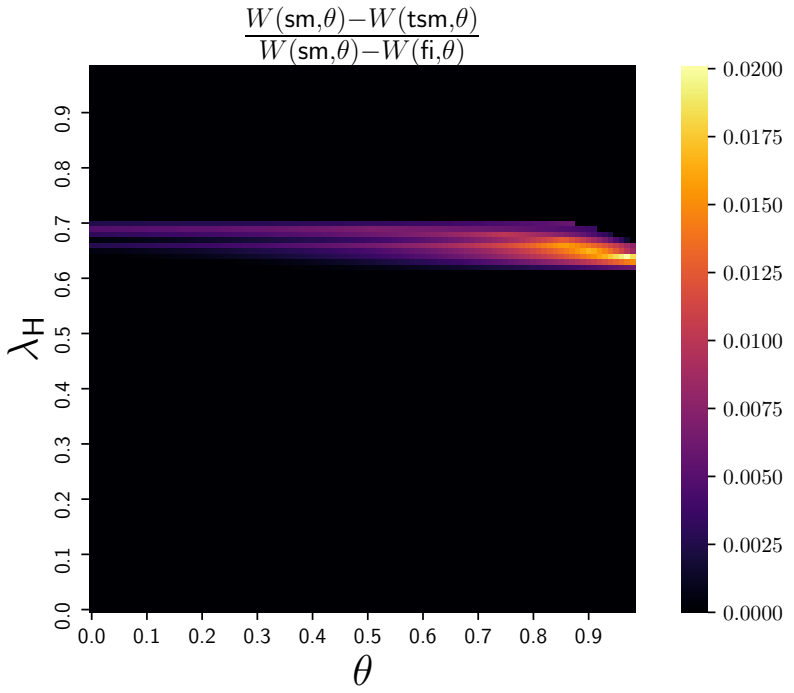


Figure EC.1 Heat map of the normalized welfare gap between optimal signaling mechanism and the best threshold mechanism (found through exhaustive search on a grid of $1/16$ increments). Model primitives: $\lambda_\mathsf{L} = 1 - \lambda_\mathsf{H}$ with $\lambda_\mathsf{H} \in [0,1]$, $\ell_\mathsf{H} = -0.7$, $u_\mathsf{L}(k) = 1 - c(k+1)$ and $u_\mathsf{H}(k) = 1 - c(k+1) - \ell_\mathsf{H}$ with $c = 0.15$.

Next, we illustrate the effectiveness of information design in a FCFS system, i.e., without a priority scheme.

In the left panel of Figure EC.2, we compare the welfare outcome of the Pareto-efficient signaling mechanism (sm) and the Pareto-efficient admission policy (ap) when we fix the service rate of H-type to be 1, but vary that of L-type from 0.8 to 1.25. (We recall that under the preemptive priority scheme, the welfare of H-type is unaffected by the signaling mechanism or the admission policy.) In particular, we plot the heat map of $W_\mathsf{L}(\mathsf{ap}) - W_\mathsf{L}(\mathsf{sm})$ on the plane $(\mu_\mathsf{L}, \lambda_\mathsf{H}) \in [0.8, 1.25] \times [0,1]$. We observe that for λ_H sufficiently large, $W_\mathsf{L}(\mathsf{ap}) = W_\mathsf{L}(\mathsf{sm})$ for any $\mu_\mathsf{L} \in [0.8, 1.25]$, implying that information design is as powerful as the first-best even when the service rate for the L-type users, μ_L, is considerably below or above its counterpart μ_H for the H-type users. To illustrate that information design remains effective even when λ_H is small, in the right panel of Figure EC.2, we plot the welfare of the L-type users under the Pareto-efficient signaling mechanism (sm) and the Pareto-efficient admission policy (ap) and the two benchmarks of full-information and no-information mechanisms when $\mu_\mathsf{L} \in [0.8, 1.25]$ and $\lambda_\mathsf{H} = 0.3$. We observe that for any μ_L, the welfare under sm remains close to that under ap and dominates that under the two benchmarks.

Next, we consider a setting where the two types differ in their service rates but not in their service priority, i.e., we revisit the FCFS queuing discipline when $\mu_\mathsf{L} \neq \mu_\mathsf{H}$. First, we remark that analyzing

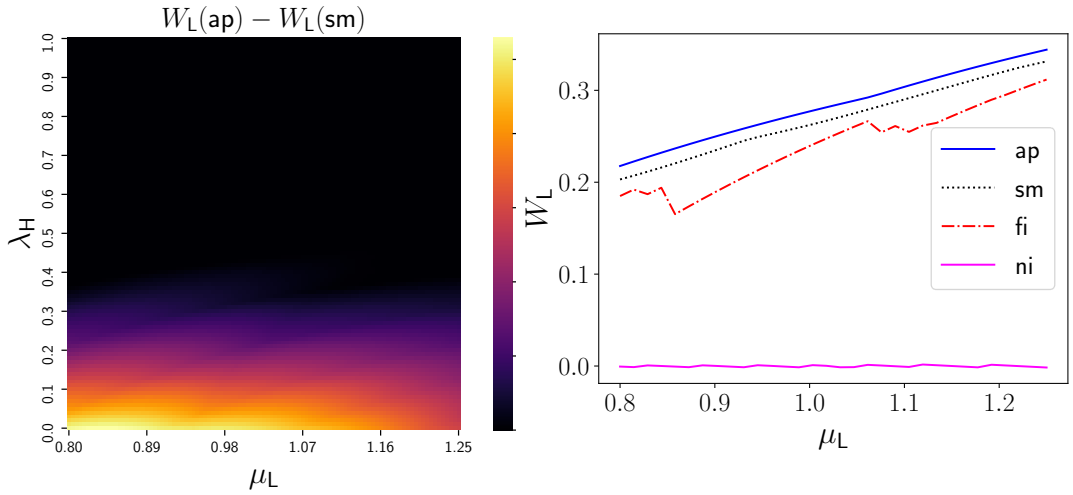


Figure EC.2 Left: Heat map of the welfare gap between the optimal admission policy and optimal public signaling mechanism, i.e., $W_{\mathsf{L}}(\mathsf{ap}) - W_{\mathsf{L}}(\mathsf{sm})$ when varying $(\mu_{\mathsf{L}}, \lambda_{\mathsf{H}}) \in [0.8, 1.25] \times [0, 1]$. Other model primitives are the same as in Figure 6. Right: Welfare outcomes for L-type under the optimal admission policy, optimal public signaling mechanism, full information, and no information when $\lambda_{\mathsf{H}} = 0.3$ and $\mu_{\mathsf{L}} \in [0.8, 1.25]$. Other model primitives are the same as in the left panel.

this setting under the FCFS service discipline is prohibitively challenging because of an explosion in the state space — it is no longer sufficient to keep track of the number of users in the queue (or even the number of users of different types). Instead, one must track the exact *sequence* of the types of users in the queue, as different type sequences (e.g., HHL vs HLH) imply different transitions in the underlying Markovian process. Because of this state space explosion, even numerically computing the Pareto efficient mechanisms under the FCFS discipline is challenging.

Nevertheless, to study the impact of information design, we restrict our attention to the class of threshold signaling mechanisms. As discussed before, this class of mechanisms are practically appealing due to their ease of implementation. To that end, we compute the Pareto-efficient threshold signaling mechanisms and compare its welfare outcomes with those of the two benchmarks of full- and no-information mechanisms as well as the welfare outcomes of the Pareto-efficient admission polices within the class of threshold policies. In Figure EC.3, we present our numerical results for the aforementioned setting and mechanisms. In particular, we consider a system with $\lambda_{\mathsf{L}} = \lambda_{\mathsf{H}} = 0.5$, $\mu_{\mathsf{H}} = 1$, and $\mu_{\mathsf{L}} \in \{1, 1.1, 1.2\}$. (The utility functions are the same as the ones described in Section 7.2.)[19] We observe that even when restricted to the class of threshold signaling mechanisms, information design results in Pareto-improvement compared to the two benchmarks of providing full or no information for all considered service rates.

[19] We note that due to lack of analytical tractability, we compute the Pareto-efficient threshold signaling mechanism and admission policies using discrete event simulations and exhaustive search over thresholds (on the expected wait

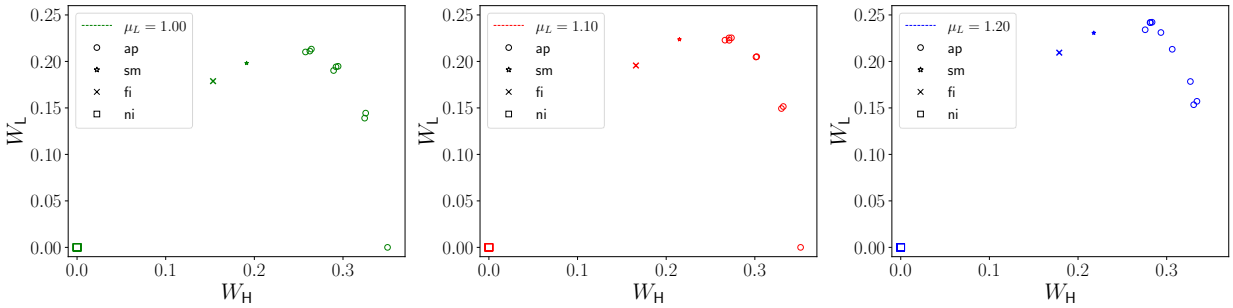


Figure EC.3 Welfare of Pareto-efficient threshold signaling mechanisms, admission policies, full-information, and no-information mechanisms for a FCFS system with $\mu_L = 1$ (left), $\mu_L = 1.1$ (center), and $\mu_L = 1.2$ (right), $\lambda_H = \lambda_L = 0.5$; Other model primitives are the same as in Figure 6.

EC.8. Exogenous Abandonment

In many situations, applicants of a social service may withdraw their request because they no longer need the service. For example, an individual seeking affordable housing may relocate to another city or move in with a partner. To include the possibility of such exogenous abandonment, in this section, we consider the same setting as introduced in Section 2, with one key modification: each arriving user has an independent deadline τ after which she no longer needs the service. More specifically, if she has not already received service by time τ after her arrival, her need for service disappears and she abandons the queue. We assume deadlines are i.i.d. and exponentially distributed with rate γ.

In the presence of exogenous abandonment, we let $u_i(k)$ denote the expected utility of a type-i user for joining when k users are already ahead in queue. Note that some of these users ahead in queue may abandon before completing their service, and the waiting time for a user is lower than than in the no-abandonment case. Furthermore, a user may obtain some utility subsequent to the abandonment. We assume all these aspects are incorporated into the utility function.

Given these modifications, we can follow the same steps as described in Section 2 and (i) establish a correspondence between signaling mechanisms and a set of all distributions satisfying obedience constraints, and (ii) characterize the Pareto frontier of the signaling mechanisms and that of the admission policies by formulating and solving linear optimization problems over feasible steady-state distributions. In particular, to obtain the Pareto frontier of signaling mechanisms, we solve the following linear program for each $\theta \in [0,1]$:

$$\begin{aligned} \max_{\pi} \quad & \theta W_L(\pi) + (1-\theta) W_H(\pi) \\ \text{subject to, } & J(\pi) \triangleq \sum_{n=0}^{\infty} \left((1+\gamma n)\pi_{n+1} - \lambda_H \pi_n \right) u_L(n) \geq 0, \end{aligned} \tag{JOIN}$$

time) with granularity of 0.1.The apparent non-convexity of the Pareto-frontier for admission policies is due to unavoidable simulation noise.

$$L(\pi) \triangleq \sum_{n=0}^{\infty} (\lambda \pi_n - (1+\gamma n)\pi_{n+1})\, u_{\mathsf{L}}(n) \le 0 \qquad \text{(LEAVE)}$$

$$\lambda_{\mathsf{H}} \pi_n \le (1+\gamma n)\pi_{n+1} \le (\lambda_{\mathsf{L}} + \lambda_{\mathsf{H}})\, \pi_n \quad \text{for all } n \ge 0 \qquad \text{(BALANCE)}$$

$$\sum_{n=0}^{\infty} \pi_n = 1, \quad \pi_n \ge 0 \quad \text{for all } n \ge 0,$$

where $W_{\mathsf{H}}(\pi)$ is as defined in (3), and $W_{\mathsf{L}}(\pi)$ is given by $W_{\mathsf{L}}(\pi) = J(\pi)$ as defined above. The main difference is in the detailed-balance conditions (BALANCE), which capture the fact that the effective arrival rate into the queue is between λ_{H} and $\lambda = \lambda_{\mathsf{L}} + \lambda_{\mathsf{H}}$, and the effective departure rate equals $1+\gamma n$ when the queue-length equals n. Furthermore, the definitions of $J(\pi)$ and $L(\pi)$ reflect the fact that the joining rate of L-type users into the queue is proportional to $(1+\gamma n)\pi_{n+1} - \lambda_{\mathsf{H}}\pi_n$ when queue-length is n, and the rate of leaving of such users is given by $\lambda\pi_n - (1+\gamma n)\pi_{n+1}$. Finally, note that the Pareto frontier of admission policies can be obtained as before by not imposing the two obedience constraints (JOIN) and (LEAVE) in the preceding program.

For our numerical analysis of this model, we continue to focus on the setting of linear utilities with the same value for service and waiting costs across the two types. It follows from a straightforward analysis that, with n users already in queue, the probability that a joining user receives service (i.e., does not abandon before being served) is given by $\frac{1}{1+(n+1)\gamma}$, and the expected time until service completion or abandonment is given by $\frac{(n+1)}{1+(n+1)\gamma}$. Taken together, the utility function of type-i users is given by $u_i(n) = \frac{1}{1+(n+1)\gamma} \cdot (1 - c(n+1)) + \frac{(n+1)\gamma}{1+(n+1)\gamma} \cdot a_i$, where a_i denotes the utility obtained by a type-i user on abandonment. Note that when $\gamma = 0$ (i.e., with no abandonment), this utility function reduces to the one considered in Section 6.

In Figure EC.4, we plot the welfare of Pareto-efficient signaling mechanisms (stars) and admission policies (circles) for different values of $\lambda_{\mathsf{L}} \in \{0.13, 0.20, 0.30\}$, with $\gamma = 0.02$, $c = 0.15$, and $a_i = 0$ for $i \in \{\mathsf{L}, \mathsf{H}\}$. Similar to the setting in Figure 2, we fix $\lambda = 1$. For each value of λ_{L}, we also plot the full-information mechanism (fi, cross) and the no-information mechanism (ni, square). We observe that for all three values of λ_{L}, the full-information and no-information mechanisms are Pareto dominated by a signaling mechanism, illustrating the power of information design over these simple information sharing benchmarks. Further, the Pareto frontier of signaling mechanisms still overlaps with that of admission policies. Taken together, this numerical example shows that our qualitative insights continue to hold in the presence of exogenous abandonment. Finally, compared with Figure 2, we observe that the welfare of both types improves as the service is less congested due to abandonment.

EC.9. General User Heterogeneity

In our baseline model, we capture the extreme of user heterogeneity by considering two user types, one of which has no viable outside option and must join the service. However, in practice, it

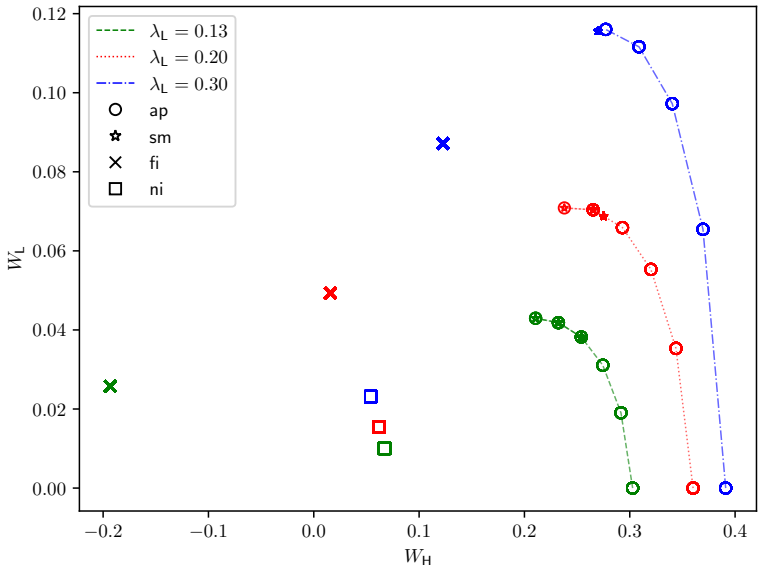


Figure EC.4 Welfare of Pareto-efficient signaling mechanisms and admission policies for $\lambda_L \in \{0.13, 0.20, 0.30\}$, $\lambda_H = 1 - \lambda_L$, $\gamma = 0.02$, $c = 0.15$, $a_L = a_H = 0$. Here, green (dashes) represents $\lambda_L = 0.13$, red (dots) represents $\lambda_L = 0.20$, and blue (dashdots) represents $\lambda_L = 0.30$. Further, circles ($\circ$) represent efficient admission policies (ap), stars ($\star$) represent efficient signaling mechanisms (sm), cross ($\times$) represents the full-information mechanism (fi), and square ($\square$) represents the no-information mechanism (ni).

is reasonable to expect a range of user types with varying levels of need for service and access to outside options. For instance, even among patients with less severe conditions who may be persuaded to avail the alternatives to an emergency room visit, the value of such alternatives might vary substantially based on the patients' symptoms. To incorporate such considerations, in this section, we extend our model to allow for multiple user types that differ in their outside options and value for service. We analyze this model numerically and show that our qualitative insights regarding the effectiveness of information design for welfare improvement continue to hold.

Suppose we have I user types, where a user of type $i \in [I]$ arrives at rate λ_i, gets utility $u_i(n)$ upon joining the queue with n users ahead of her, and has an outside option of $\ell_i \in \mathbb{R} \cup \{-\infty\}$. Here, $\ell_i = -\infty$ captures the case where type-i users have no viable outside option. (We assume no abandonment in this section.) Our baseline model corresponds to the case $I = 2$ with $\ell_1 = 0$ and $\ell_2 = -\infty$.

In practice, a social service provider may not always be able to observe the type of a user. Moreover, ethical concerns may limit a service provider from making information provision depend on the users' outside options. Such limitations may make private signaling infeasible, and due to such practical considerations, we focus on *public signaling mechanisms*. Note that in our baseline

model, public and private signaling are the same because high-need users have no outside option and always join irrespective of the belief.

For public signals, using the revelation principle, one can show that it suffices to consider signaling mechanisms where signals correspond to subsets $S \subseteq [I]$, and which are obedient in the sense that when the signal is $S \subseteq [I]$ only users with type $i \in S$ find it optimal to join the queue, whereas users with type $j \notin S$ find it optimal to leave. Focusing on such signaling mechanisms, similar arguments as in Section 2 allow us to formulate a linear program to compute the Pareto-efficient (public) signaling mechanisms. To see this, for a public signaling mechanism, let $x_{n,S}$ denote the joint probability (in steady-state) that the queue-length upon arrival of a user is n and the user receives the signal $S \subseteq [I]$, and note that $\sum_{S\subseteq[I]} x_{n,S}$ denotes the probability that the queue-length is n in steady-state. The detailed-balance condition can then be written as $\sum_{S\subseteq[I]} \lambda_S x_{n,S} = \sum_{S\subseteq[I]} x_{n+1,S}$, where $\lambda_S = \sum_{i\in S} \lambda_i$ denotes the total arrival rate of the users with type $i \in S$. The welfare of type-i users can then be written as a function of $x = \{x_{n,S} : n \geq 0, S \subseteq [I]\}$ as follows:

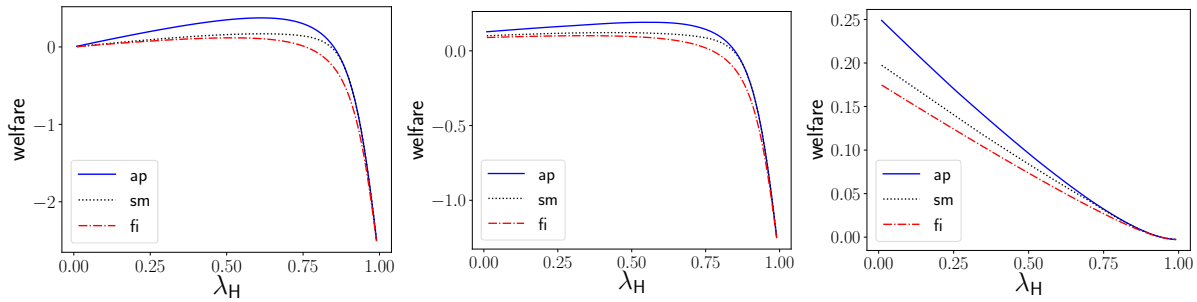
$$W_i(x) = \lambda_i \left(\sum_{n\geq 0} \sum_{S \ni i} x_{n,S} u_i(n) + \sum_{n\geq 0} \sum_{S \not\ni i} x_{n,S} \ell_i \right).$$

Further, upon receiving a signal $S \subseteq [I]$, since a user with type $i \in S$ finds it optimal to join the queue, this implies $\sum_{n\geq 0} x_{n,S}(u_i(n) - \ell_i) \geq 0$ for $i \in S$. Similarly, since a user with type $i \notin S$ finds it optimal to leave, we have $\sum_{n\geq 0} x_{n,S}(u_i(n) - \ell_i) \leq 0$ for $i \notin S$. Putting it all together, it follows that the Pareto-efficient public signaling mechanisms correspond to the optimal solutions of the following linear program for different choices of non-negative weights $\theta = (\theta_i : i \in [I])$:

$$\begin{aligned}
\max_{x} \quad & \sum_{i\in[I]} \theta_i W_i(x) \\
\text{subject to,} \quad & \sum_{n\geq 0} x_{n,S}(u_i(n) - \ell_i) \geq 0, \quad \text{for } i \in S \text{ and } S \subseteq [I], \\
& \sum_{n\geq 0} x_{n,S}(u_i(n) - \ell_i) \leq 0, \quad \text{for } i \notin S \text{ and } S \subseteq [I], \\
& \sum_{S\subseteq[I]} \lambda_S x_{n,S} = \sum_{S\subseteq[I]} x_{n+1,S}, \quad \text{for all } n \geq 0, \\
& \sum_{n\geq 0} \sum_{S\subseteq[I]} x_{n,S} = 1, \text{ and } x_{n,S} \geq 0 \ \text{ for all } n \geq 0, \text{ and } S \subseteq [I].
\end{aligned}$$

Using this linear program, we numerically investigate the effectiveness of information design by comparing it to the full-information mechanism and the Pareto-efficient admission polices.[20] In particular, we consider an example with three users types, all with the same linear utility function

[20] We compute the Pareto-efficient admission polices for any given weight $\theta = (\theta_i : i \in [I])$ by dropping the obedience constraints from the preceding linear program.



(a) $\theta = (0,0,1)$ (b) $\theta = (\frac{1}{4},\frac{1}{4},\frac{1}{2})$ (c) $\theta = (\frac{1}{2},\frac{1}{2},0)$

Figure EC.5 Welfare of the Pareto-efficient (public) signaling mechanism $\mathsf{sm}(\theta)$, the Pareto-efficient admission policy $\mathsf{ap}(\theta)$, and the full-information mechanism fi for three types $I = \{1,2,3\}$. Here, the arrival rates are given by $(\lambda_1,\lambda_2,\lambda_3) = (\lambda_\mathsf{L},\lambda_\mathsf{L},\lambda_\mathsf{H})$ with $\lambda_\mathsf{L} = (1-\lambda_\mathsf{H})/2$. The outside options equal $(\ell_1,\ell_2,\ell_3) = (0,-0.25,-\infty)$, and $u_i(k) = 1-c(k+1)$ with $c = 0.15$.

$u(n) = 1-c(n+1)$ with $c = 0.15$ but different outside options given by $(\ell_1,\ell_2,\ell_3) = (0,-0.25,-\infty)$. In particular, the first two types have viable outside options, while the third type has no viable outside option. In keeping with the terminology of our baseline model for easier comparison, we refer to the third-type as the H type, and the first two types as L types. The arrival rates are given by $(\lambda_1,\lambda_2,\lambda_3) = (\lambda_\mathsf{L},\lambda_\mathsf{L},\lambda_\mathsf{H})$ with $\lambda_\mathsf{L} = (1-\lambda_\mathsf{H})/2$.

In Figure EC.5, we plot the welfare $W(\pi,\theta)$ as a function of the arrival rate λ_H for the full-information mechanism, the optimal (public) signaling mechanism $\mathsf{sm}(\theta)$ and the optimal admission policy $\mathsf{ap}(\theta)$ for $\theta = (\theta_1,\theta_2,\theta_3) \in \left\{(0,0,1),\left(\frac{1}{4},\frac{1}{4},\frac{1}{2}\right),\left(\frac{1}{2},\frac{1}{2},0\right)\right\}$. Similar to our observations in Figure 3, we observe that information design results in welfare improvement over the full-information mechanism when the user population is fairly balanced (given our parametrization of the arrival rates, this corresponds to λ_H being not too large). Further, for large enough λ_H, the optimal signaling mechanism $\mathsf{sm}(\theta)$ achieves the same welfare as that of the optimal admission policy $\mathsf{ap}(\theta)$. This point is further illustrated in Figure EC.6 (left panel), where, for the weight parametrization $\theta = (\frac{\theta_\mathsf{L}}{2},\frac{\theta_\mathsf{L}}{2},1-\theta_\mathsf{L})$, we display the region in the $(\theta_\mathsf{L},\lambda_\mathsf{H})$ plane where the optimal signaling mechanism achieves the same welfare as the optimal admission policy. Finally, Figure EC.6 (right panel) shows that even in regions where the two policies do not achieve the same welfare, the welfare gap is fairly small.

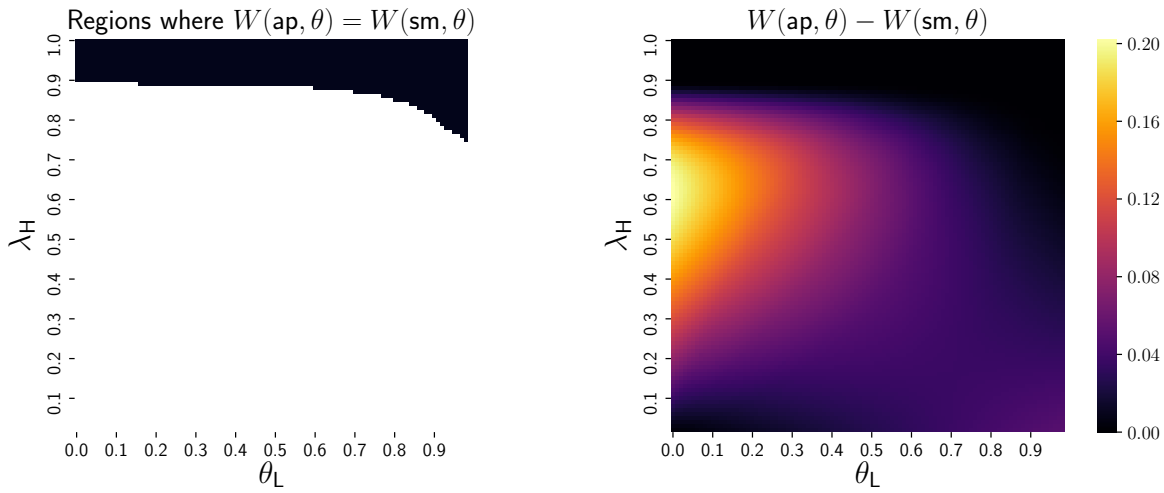


Figure EC.6 **Left: Regions of the $(\theta, \lambda_\mathsf{H})$ plane for which $\mathsf{sm}(\theta) = \mathsf{ap}(\theta)$, i.e., the signaling mechanism $\mathsf{sm}(\theta)$ is Pareto-efficient within Π_{AP}. Right: Heat map of the welfare gap between the optimal admission policy and optimal public signaling mechanism, i.e., $W(\mathsf{ap},\theta) - W(\mathsf{sm},\theta)$. Model primitives: The arrival rates are given by $(\lambda_1, \lambda_2, \lambda_3) = (\lambda_\mathsf{L}, \lambda_\mathsf{L}, \lambda_\mathsf{H})$ with $\lambda_\mathsf{L} = (1 - \lambda_\mathsf{H})/2$. The outside options equal $(\ell_1, \ell_2, \ell_3) = (0, -0.25, -\infty)$, $u_i(k) = 1 - c(k+1)$ with $c = 0.15$, and the welfare weights are given by $\theta = (\theta_\mathsf{L}/2, \theta_\mathsf{L}/2, 1 - \theta_\mathsf{L})$.**