On the Efficiency of Queueing in Dynamic Matching Markets*

Laura Doval† Balázs Szentes‡

December 5, 2024

Abstract

We study a two-sided dynamic matching market where agents arrive randomly. An arriving agent is immediately matched if agents are waiting on the other side. Otherwise, the agent decides whether to exit the market or join a queue to wait for a match. Waiting is costly: agents discount the future and incur costs while they wait. We characterize the equilibrium and socially optimal queue sizes under first-come, first-served. Depending on the model parameters, equilibrium queues can be shorter or longer than efficiency would require them to be. Indeed, socially optimal queues may be unbounded, even if equilibrium queues are not. By contrast, when agents only incur flow costs while they wait, equilibrium queues are typically longer than socially optimal ones (cf. Baccara et al., 2020). Unlike one-sided markets, the comparison between equilibrium and socially optimal queues in two-sided markets depends on agents' time preferences.

Keywords: *dynamic matching, queueing, two-sided markets, efficiency, discounting, flow costs*
JEL codes: C61, C78, D47

1 Introduction

Forum (2019)In many matching markets, agents arrive over time and matches are created sequentially. Examples of such dynamic markets include one-sided markets, such as the allocation of deceased-donor organs and public housing, as well as two-sided markets, such as kidney exchange, child adoption, and platforms for barter exchange. A key determinant of the allocation in these markets is agents' patience. If agents are willing to wait more, the quality of matches might improve at the cost of increased waiting times. We contribute to this problem in a small way: we analyze equilibria in a simple two-sided matching market and compare the resulting

*We thank the Editor, Sushil Bikhchandani, the Advisory Editor, and two anonymous referees for feedback that has greatly improved this paper. We thank Yangfan Zhou for exceptional research assistance.
†Columbia Business School and CEPR. E-mail:laura.doval@columbia.edu
‡HKU Business School and London School of Economics. E-mail: szentes@hku.hk

allocation with the socially efficient one. We fully characterize when the agents are more or less patient than efficiency would require them to be, and we show how our conclusions depend on agents' time preferences.

In a two-sided market, an agent's decision whether to join a queue has opposing effects on social welfare. On the one hand, joining the queue increases the waiting time for subsequent agents on her side of the market. Hence, an agent's decision to wait decreases the welfare of others on the same side of the market. On the other hand, joining the queue creates a positive externality for agents on the opposite side by reducing their waiting times and potentially increasing the number of matches created. When deciding whether to join a queue, an agent ignores *all* the external effects of her decision. The goal of this paper is to understand what factors determine the relative magnitude of these two effects by comparing equilibrium queues with efficient ones. Our results show that a key factor is how agents' time preferences are specified.

In our model, agents arrive stochastically on each side of the market according to a Poisson process with arrival rate λ. If there is a queue in front of an arriving agent, she has to decide whether to join the queue or leave the market and take her outside option. Throughout, we assume that the queueing discipline is first-come first-served (henceforth, FCFS).[1] An agent's payoff from matching, H, exceeds the value of her outside option, L, but waiting to be matched is costly. Each agent discounts the future by rate r and pays a flow cost c each moment she waits in the queue. Our model thus encompasses the polar cases of *pure discounting*—$r > 0, c = 0$—and *pure flow costs*—$r = 0, c > 0$—which are the leading specifications in economics and operations research, respectively.[2] The cases of pure discounting and pure flow costs also allow us to capture in reduced form different allocation mechanisms such as waiting lists—in which agents incur no costs while waiting, but may discount the future—and (pure) queueing—in which agents incur costs while waiting (cf. Elster, 1992).[3]

To analyze which of the two opposing effects dominates in our two-sided environment, we compare equilibrium queueing behavior with the socially optimal one. The latter maximizes an agent's steady-state payoff. We show an agent's equilibrium behavior is defined by a threshold, k^*. That is, an agent joins a queue if and only if the queue on her side of the market is shorter than k^*. Similarly, the socially optimal policy is characterized by the length of the longest queue the social planner wishes to maintain, which we refer to as the queue size and denote by K^*. Notably, both the equilibrium and socially optimal thresholds depend on two parameters. The first is the *normalized outside option*, defined as the ratio of the payoffs L and H, modified by the discounted cost of waiting indefinitely, $(L + c/r)/(H + c/r)$. The normalized outside

[1] We focus on FCFS because of its pervasiveness in practice. We do not intend to characterize the optimal queueing discipline.

[2] In his treatise on the perceived fairness of FCFS, Larson (1987) argues that "the actual and/or perceived utility of participating in the system is (1) a nonlinear function of queueing delay".

[3] Lindsay and Feigenbaum (1984) propose discounting as the mechanism through which waiting lists, which allow individuals to dispose of their time freely while they wait, may be used for rationing purposes.

option captures the relative benefit of joining the queue.[4] The second is the *discounted arrival rate*, $\lambda/(\lambda + r)$, which captures the average discounted time an agent waits for an arrival on the opposite side. To understand whether in equilibrium agents form inefficiently long queues or are not willing to wait enough, we compare k^* and K^*.

Section 4 presents our main results. First, we characterize the parameter values for which the socially optimal queue size K^* is longer or shorter than the equilibrium queue size k^* (see Propositions 1 and 2). In particular, we show that when the discounted arrival rate is small or the normalized outside option is high, the socially optimal queue is longer than the equilibrium queue. By contrast, when the discounted arrival rate is large relative to the normalized outside option, the equilibrium queue is longer than the socially optimal one. Figure 1a illustrates the comparison between K^* and k^* based on computations in the case of pure discounting.[5] Finally, we show that for a certain set of parameters, the socially optimal queue is unbounded (i.e., $K^* = \infty$), whereas the equilibrium queue is always bounded (see Proposition 3). In fact, the socially optimal queue can be unbounded even if agents are not willing to queue at all.

That the planner may wish to implement longer queues than equilibrium ones contrasts both with the conclusions of the seminal work of Naor (1969) on customer strategic behavior in a server model and those of Baccara et al. (2020) on two-sided markets. Indeed, both papers conclude equilibrium queues are longer than socially optimal ones. To better compare our results with those in the literature, we specialize our model to the case of pure flow costs, in which the agents do not discount the future ($r \approx 0$), but still incur a cost c while waiting. We show in Proposition 4 the equilibrium queue size is typically (weakly) larger than the socially optimal one in the pure flow costs case.[6] In other words, when agents do not discount the future, the negative externality an agent imposes on her own side of the market generally dominates the positive externality she imposes on the other side. Based on computations, Figure 1b depicts the comparison between K^* and k^* in the pure flow costs case as a function of the difference in values between matching in the queue and taking the outside option, $H - L$, and the cost-adjusted arrival rate, λ/c.

We now illustrate one of the mechanisms through which discounting affects the comparison between socially efficient and equilibrium queue sizes by considering why unbounded queues are possible under discounting; we leave the intuition for the rest of the results to the main text. To do so, consider the following example. Suppose $H = 3$, $L = 1$, agents are very impatient, $r \approx \infty$, incur no flow costs, $c \approx 0$, and agents arrive infrequently, $\lambda \approx 0$. Because agents are infinitely impatient and incur no flow costs, their payoff is zero whenever they wait. Therefore, in equilibrium, agents never wait ($k^* = 0$) and their payoff is the value of the outside option, 1. Furthermore, note an agent's steady-state payoff cannot exceed 2. The reason is that in

[4]This is immediate when $c = 0$, in which case $(L + c/r)/(H + c/r)$ simplifies to L/H. Remark 2 explains why the intuition in the main text still holds when $c > 0$.

[5]Figure 1a and all figures in the paper are constructed by numerically evaluating the model in Mathematica and exporting the output to TEX. Because queue size is an integer, some figures display a sawtooth pattern.

[6]In fact, the only exception is when $K^* = 1$, in which case, it is possible that $k^* = 0$.

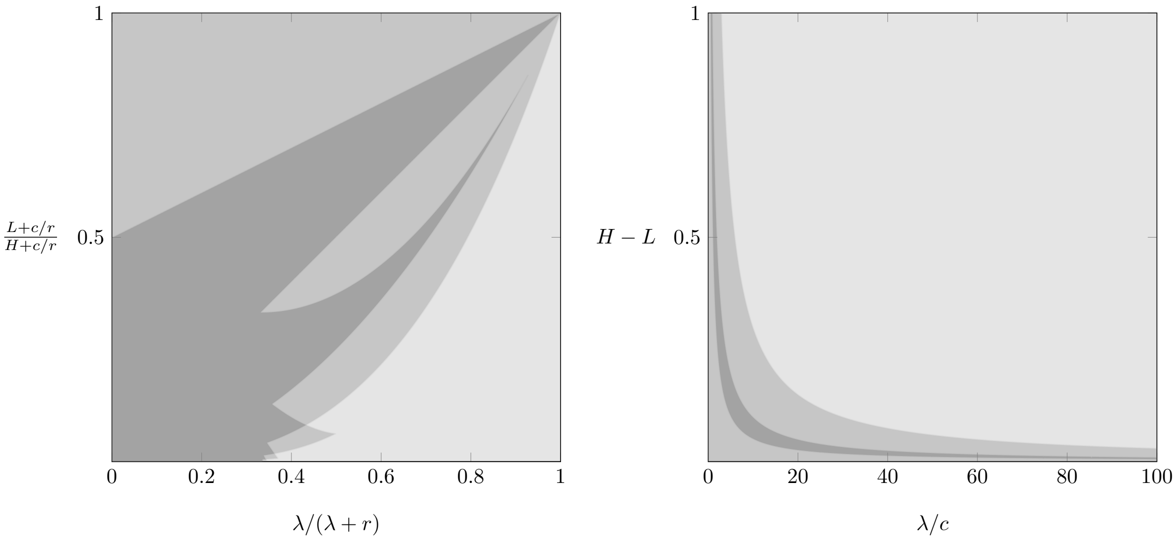

Figure (a) Discounting ($r > 0, c \geq 0$)

Figure (b) Pure flow costs ($r = 0, c > 0$)

Figure 1: Comparison between K^* and k^*: $K^* > k^*$ (dark gray), $K^* = k^*$ (gray), $K^* < k^*$ (light gray)

each created match one of the agents has to wait and her payoff is zero, whereas the other is matched immediately upon arrival and receives a payoff of 3. So, agents who are matched receive a payoff of 1.5, on average, and the payoff of those agents who take their outside option is 1. This argument also implies agents' average payoff is maximized if each agent is matched. Therefore, the socially optimal policy is to make an agent join a queue regardless of how long that queue is, that is, $K^* = \infty$. In summary, when agents discount their future and incur no flow costs, the payoff of each agent is bounded by zero, and hence, the opportunity cost of forcing an agent to join a queue never exceeds the outside option, L. Therefore, as long as the benefit from making an agent wait, H, is sufficiently large, socially optimal queues can be long. By contrast, when agents pay a flow cost of waiting *and* do not discount the future, their payoffs are unbounded from below. Therefore, forcing an agent to join a sufficiently long queue would never be optimal.

Finally, Section 5 considers two extensions of our model that break the symmetry across sides. Section 5.1 considers the case in which each side has different outside options, and Section 5.2 considers the case in which each side has different arrival rates. In both cases, the socially optimal policy specifies a possibly different queue size for each side. When outside options differ across sides, we show the socially optimal queue sizes are still symmetric (Proposition 5) and that equilibrium queue sizes continue to be typically longer than socially optimal ones in the pure flow costs case (Proposition 6). When arrival rates differ across sides, we provide sufficient conditions under which the socially optimal queue size for the slow-arriving side is unbounded, and we characterize the optimal queue size for the fast-arriving side (Proposition 7).

To summarize, we make two contributions. First, we characterize the comparison between equilibrium and socially efficient queue sizes in a two-sided market, in which agents discount the future and incur flow costs while they wait. We show that, depending on parameter values, equilibrium queues may be shorter or longer than socially optimal ones. Our model can thus provide a unified lens through which to understand why some markets, like child adoption, are considered congested, while others, like barter exchanges, are viewed as thin (see Baccara et al. (2020), Prendergast and Stole (2001), and the references therein). In the former, agents have few outside options, leading to equilibrium queues that are longer than socially optimal. In the latter, because agents have good alternatives to barter exchange in a fully monetized economy, equilibrium queues are shorter than socially optimal. Second, in contrast to one-sided markets, we show the comparison between socially optimal and equilibrium queue sizes depends on the specification of agents' time preferences in two-sided markets. In particular, equilibrium queues are typically longer than socially optimal ones under pure flow costs.

Together, our results imply that to implement efficient queues, the planner may need a richer set of policies in two-sided markets with discounting than in two-sided markets with pure flow costs or one-sided markets. Whereas in the latter efficiency can be achieved via admission control policies, such as tolls or caps on queue length, in two-sided markets with discounting, the planner may also need to incentivize entry, for instance, through subsidies. Whether agents discount the future, or arrival rates and outside options are such that socially efficient queues are shorter than equilibrium ones, is an empirical question. At the same time, pure time-discounting is the leading preference specification in empirical market design studies of assignment through queueing or waiting lists (see, e.g., Agarwal et al. (2021) and Waldinger (2021) for deceased-donor organs and public housing, respectively, and Huitfeldt et al. (2024) for the assignment of patients to physicians). When considering two-sided applications, the analysis herein can then inform researchers about the effects of counterfactual designs as a function of the estimated parameters, which may not be available under pure flow costs.

Related Literature The paper contributes to the literatures on equilibrium behavior of customers in queueing systems and dynamic matching. With exceptions (see below), the analysis of strategic behavior in queueing systems is mainly concerned with server models, that is, one-sided markets in which customers choose whether to exit and take an outside option, or to join a queue to get serviced by a server. If the queueing discipline is FCFS, the only external effect of an agent's decision to join a queue is that she increases the waiting time of others. Consequently, equilibrium queues are inefficiently long, as was first noted by Naor (1969). Hassin (1985) argues a last-come-first-served queueing discipline implements the optimal allocation in this setting. Our paper extends the classic models in this literature to allow for two sides, but we restrict attention to the commonly used FCFS discipline. This focus allows us to study the aforementioned positive externality not present in the previous papers on the equilibrium behavior: when a customer joins a queue, she exerts a positive externality on the customers arriving on the other side, by creating matching opportunities.

Among the papers on the recent literature on dynamic matching, our analysis is closest to that in Baccara et al. (2020). The authors analyze a two-sided dynamic matching market in which agents have a binary quality. An agent's payoff depends on her own quality and that of her match and it is supermodular. A key result in Baccara et al. (2020) is that low-quality agents are willing to match with anybody, so the equilibrium queue length is determined by the high-quality agents' decision of whether to wait or match with low-quality agents. Baccara et al. (2020) show the negative externality caused by an agent on her side of the market always dominates the positive externality imposed on the other side. Consequently, equilibrium queues are inefficiently long in their model. In Section 4.1, we argue that when specialized to the case of pure flow costs, our model is similar to theirs, with the outside option playing the role of the low-quality agents in their model. As mentioned above, in that section, we confirm the result of Baccara et al. (2020). We also explain why, under discounting, socially optimal queue sizes may be longer than equilibrium ones, whereas in the pure flow costs case, they typically are not.

Ünver (2010), Akbarpour et al. (2020), Anderson et al. (2015), and Ashlagi et al. (2019) consider dynamic matching models in which preferences are compatibility-based: agents get a payoff of 0 when unmatched, and 1 when matched with a compatible agent. Thus, in these models, agents are always more impatient than the planner, and these papers focus on understanding whether forcing agents to wait to generate more matches in the future is optimal for the planner. Leshno (2022) considers a model in which agents and objects, of one of two types each, arrive over time. The planner's objective is to minimize the number of agents who exit the system matched to an object of the opposite type, whereas agents may prefer to take this object if getting the object that matches their type takes too long. As in the previous papers, agents in Leshno (2022) are always more impatient than the planner.

Recent papers consider more general time preference specifications or objectives in the context of one-sided markets. Focusing on the welfare of the agents in the queue, Schummer (2021) shows FCFS Pareto dominates the outcome of a general class of policies in the pure flow costs case or when agents are risk averse, but not when agents discount the future. Because the model and welfare criteria are different, the reason discounting reverses the results relative to pure flow costs in Schummer (2021) differs from that in our model. Under pure flow costs, Che and Tercieux (2021) show FCFS maximizes a weighted sum of agents' and service provider's payoff when agents can renege the queue and the designer can design both the queueing discipline and agents' information. Allowing for agents with heterogeneous lifespans, Nikzad and Strack (2024) show that service in random order is the most equitable queueing discipline among those that prioritize agents with longer waiting times, and if lifetimes are exponentially distributed, it maximizes utilitarian welfare under a broad class of time preferences, which includes time-discounting.

Formally, our model is an instance of a double-ended queue, which Kendall (1951) introduced to model the matching of passengers to taxis. Dobbie (1961) introduced solution methods. Motivated by applications, most papers in this literature study performance measures under FCFS (see, e.g., Zenios (1999) and Boxma et al. (2011) for organ waiting lists, Afèche et al. (2014) and Diamant and Baron (2019) for financial markets, Adan and Weiss (2012) for public housing, and Gurvich and Ward (2015) for assemble-to-order manufacture). Motivated by ride-sharing applications, a recent literature studies passenger strategic behavior for a fixed and finite taxi queue size, sometimes contrasting it with the socially optimal passenger joining strategy. Shi and Lian (2016), Wang et al. (2017), and Wang et al. (2023) study both passengers' optimal strategies and the optimal taxi queue size given passengers' strategic behavior. All these papers restrict attention to the pure flow costs case. Moreover, because they do not consider the taxi driver's behavior, they are not concerned with the positive externality that passengers impose on taxi drivers when they join the queue. Nguyen and Phung-Duc (2022a,b) consider the strategic behavior of both sides, but do not consider the socially optimal policy.

Finally, some of the ideas we present have antecedents in the literature of matching with frictions. First, that in two sided economies agents may wait too long to match (failing to internalize the negative externality on their own side), or too little (failing to internalize the positive externality on the other side) already appears in Shimer and Smith (2001). In their model, ex ante heterogeneous agents have to (costly) search to find a partner on the other side. They show that to decentralize the efficient solution, some agents have to be subsidized (because they search too little), whereas others have to be taxed (because they search too much). Second, that in two-sided economies the predictions under discounting and flow costs of waiting differ is also made by Atakan (2006) in the context of a model of search. Because discounting alters the marginal value of waiting for a better match, whereas flow costs do not, Atakan (2006) finds assortative matching holds in a frictional search model with pure flow costs under weaker conditions than with discounting (cf. Shimer and Smith, 2000).

2 Model

We study a two-sided market in continuous time. On each side, agents arrive according to a Poisson process with arrival rate λ. Agents are matched on a first-come, first-served basis. Concretely, if there is a queue on one side of the market and an agent arrives on the other side, the agent who waited the longest in the queue is matched with the arriving agent and they leave the market. An arriving agent who is not matched immediately has to decide whether to join the (possibly empty) queue or leave the market. If the agent joins the queue, she remains there until she matches with an agent on the other side, which yields a payoff of H; if she leaves the market unmatched, she receives a payoff of $L < H$.[7] Agents discount the future according to a

[7] Taking the outside option can capture in reduced-form (i) matching with lower quality agents, as when we connect our result with Baccara et al. (2020) in Section 4.1, (ii) matching with a less preferred object, as in Leshno

common discount rate r, and every period they are in the queue, agents pay a flow cost $c > 0$.

Below, the following two notation conventions are useful. First, because both the equilibrium and socially optimal queue sizes depend on the discounted arrival rate, $\lambda/(\lambda + r)$, it is useful to denote this quantity by x, that is, $x \equiv \lambda/(\lambda + r)$. The discounted arrival rate is the average discounted time until the next arrival, capturing that agents adjust the arrival rate λ by the discount rate r.[8] Second, both the equilibrium and socially optimal queue sizes depend on the payoffs H and L modified by the discounted cost of waiting indefinitely in the queue, c/r. Hence, we denote the quantities $H + c/r$ and $L + c/r$ by H_c and L_c, respectively.

2.1 Equilibrium queue

An agent's decision whether to join the queue or leave the market depends only on the length of the queue she faces. Let $u(k)$ denote the payoff of an agent who joins a queue of length $k - 1$. The value function u is recursively defined by the following equation:

$$ru(k) = \begin{cases} \lambda(H - u(1)) - c & \text{if } k = 1 \\ \lambda(u(k-1) - u(k)) - c & \text{if } k > 1 \end{cases}. \tag{1}$$

The first line states the dividend from forming a queue of size one, $ru(1)$, is the product of the arrival rate of a match, λ, and the increase in value from $u(1)$ to H *minus* the cost of waiting in the queue, c. The second line states the dividend from joining a queue of size $k - 1$ is again the product of the arrival rate λ and the increase in value from jumping ahead in the queue by one position, $u(k-1) - u(k)$, minus the cost of waiting in the queue.

Solving for u by induction yields

$$u(k) = x^k H - (1 - x^k)\frac{c}{r}. \tag{2}$$

In words, an agent's payoff from joining the queue in position k, $u(k)$, is her discounted payoff from matching, $x^k H$, minus her discounted flow costs.

Consequently, an agent who arrives to a queue of length $k - 1$ on her side is willing to join the queue if and only if

$$x^k H - (1 - x^k)\frac{c}{r} \geq L \Leftrightarrow x^k H_c \geq L_c,$$

where recall that $H_c = H + c/r$ and $L_c = L + c/r$. It follows that the longest queue formed in

(2022), (iii) matching in a competing exchange with a lower service rate (cf. Das et al., 2015), or (iv) in the context of kidney exchange, choosing to participate in list kidney exchange. Regarding (i) and (ii), our one-type model can encompass the two-type formulations in Baccara et al. (2020) and Leshno (2022) because their models can be analyzed as a one-dimensional Markov chain; see Che and Tercieux (2021) and Doval (Forthcoming) for a formal argument and derivations.

[8] Formally, when t is Poisson distributed with rate λ, $\mathbb{E}\left[\exp^{-rt}\right] = x$.

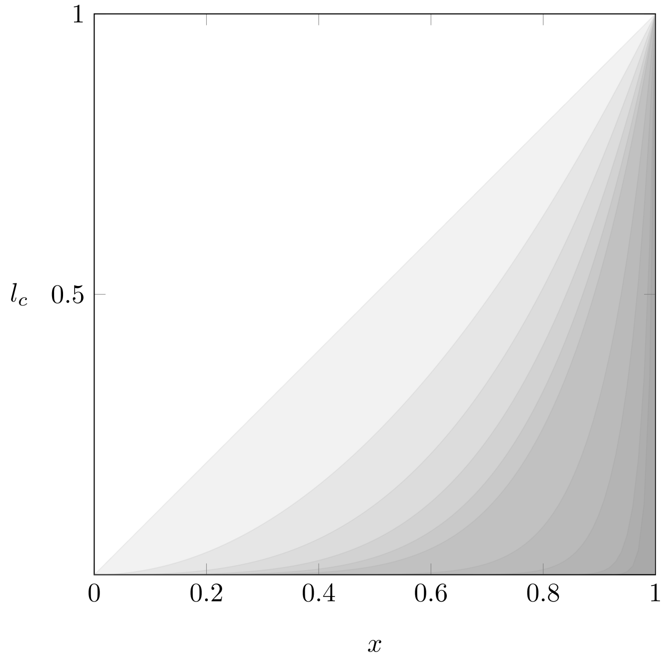

Figure 2: Equilibrium queue size, k^*. The white area corresponds to $k^* = 0$. The increasingly darker shades of gray depict $k^* \in \{1, \dots, 5, [6, 10], [11, 20], [21, 50], [51, 99], [100, +\infty)\}$.

equilibrium, k^*, is defined by

$$x^{k^*} \geq \frac{L_c}{H_c} > x^{k^*+1}.$$

We refer to the ratio L_c/H_c as the normalized outside option and denote it by l_c. The equilibrium queue size is given by

$$k^* = \left\lfloor \frac{\log l_c}{\log x} \right\rfloor. \tag{3}$$

Figure 2 depicts the values of k^* as a function of x and l_c, where the white area corresponds to $k^* = 0$, and darker shades of gray correspond to higher values of k^*. From Equation 3 we can conclude the following intuitive comparative static result:

Remark 1 (Comparative statics for k^*)**.** *The largest equilibrium queue size, k^*, is decreasing in l_c and r, and increasing in λ.*

Remark 1 says an agent's willingness to join the queue is higher if either her outside option or her waiting costs are smaller (l_c is small), she is more patient (r is small), or if agents on the other side of the market arrive more frequently (λ is high).

An argument attributed to Bell (1971) explains why the flow cost c enters only through the modified payoffs H_c and L_c; we include it in Remark 2 and it may be skipped with little loss of continuity:

Remark 2. *Consider an alternative model in which agents discount the future at rate r and experience no flow costs (i.e., $c = 0$). Instead, to join the queue, an agent must pay c/r upfront and receives a refund of c/r upon being matched. In other words, the opportunity cost of joining the queue is $L + c/r = L_c$, and upon being matched, the agent's payoff is $H + c/r = H_c$.*

It is immediate to show that in this alternative model, the equilibrium queue size is also given by k^* as in Equation 3. As the analysis that follows shows, the socially optimal queue size is also the same across the models. For this reason, below, we sometimes refer to L_c and H_c as the values of the outside option and of being matched, respectively, when doing so is not likely to lead to confusion.

3 Socially optimal queue size under FCFS

In this section, we analyze the socially optimal queue size under FCFS, which we then compare with the equilibrium one in Section 4. We consider the problem of a social planner who can force an agent on one side to wait for agents on the other side to arrive (queue), but can also prevent them from doing so, forcing them to take their outside option. We assume the social planner's goal is to maximize the steady state welfare of a randomly arriving agent.[9] We focus on this Pareto welfare criterion for two reasons. First, under this criterion, the social planner weighs all generations of agents equally. Instead, assuming the social planner maximizes the discounted sum of utilities implies the planner favors the welfare of current generations over that of future ones. Thus, under the latter criterion, the socially optimal queueing policy would balance both the agents' and the social planner's time preferences. Second, this welfare criterion allows for a *ceteris paribus* comparison of the socially optimal queues across the different models of agents' time preferences we consider.

We restrict the planner to choosing the queue size; that is, the planner's policy is characterized by a cutoff k such that the planner makes an agent queue if and only if there are less than k agents on the side of the arriving agent already in the market, and there is nobody on the other side. In what follows, we denote by $\mathrm{W}(k)$ the value of welfare if the cutoff is k. In other words, $\mathrm{W}(k)$ is the steady-state payoff of an arriving agent if the planner's cutoff is k.

[9] One can easily show this social planner has limits-of-the-mean preferences. More formally, the planner's policy maximizes

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{N} u_i}{N},$$

where u_i denotes the expected payoff of the ith arriving agent.

The socially optimal queue size, K^*, is the solution to

$$\max_{k\in\mathbb{N}_0} \mathrm{W}(k).$$

The Markov Process The steady-state payoff of an arriving agent is determined by the distribution of queue lengths. Each planner's cutoff, k, and the arrival process induce a Markov process and an ergodic distribution over the possible queue lengths, $\{0, 1, ..., k\}$. Let π_s denote the ergodic probability that the length of the queue on a given side of the market is s. We do not introduce notation indicating the side of the market on which the queue is formed, but we note that, by symmetry, these probabilities are the same on the two sides. Note agents arrive on either side with equal probabilities.

Consider an interval of time of length Δ. The following three heuristic equations establish the relationship between the ergodic probabilities of different states:

$$\begin{aligned}
\pi_0 &= (1-2\lambda\Delta)\pi_0 + 2\lambda\Delta\pi_1, \\
\pi_s &= (1-2\lambda\Delta)\pi_s + \lambda\Delta\pi_{s-1} + \lambda\Delta\pi_{s+1}, s \in \{1, ..., k-1\}, \\
\pi_k &= (1-\lambda\Delta)\pi_k + \lambda\Delta\pi_{k-1}.
\end{aligned} \tag{4}$$

The first line in Equation 4 states that the probability the queue length is 0 after time Δ is the sum of the probability that the initial queue length is 0 and no one arrives on either side, and the probability that the initial queue length is 1 and an agent arrives on the opposite side.[10] The second line states that for $s \in \{1, ..., k-1\}$, the probability the queue length is s after Δ time is the sum of the probabilities of three events: (i) s agents are queueing initially and no one arrives on either side, (ii) $s-1$ agents are queueing on one side and an additional agent arrives on that side, and (iii) $s+1$ agents are queueing and an agent arrives on the opposite side. Finally, the probability the queue length is k is the sum of the probability that (i) k same side agents are queueing and either no one arrives on either side, or someone arrives on that same side and exits, and (ii) $k-1$ agents are initially queueing on one side, and an agent on that side arrives.

A well-known result in the theory of Markov chains implies that because the Markov process is irreducible and aperiodic and its transition matrix is doubly-stochastic, the (unique) steady-state distribution is uniform (see Ross (2010, Chapter 4, Exercise 20)).[11] That is,

$$\pi_s = \frac{1}{2k+1}, \tag{5}$$

for each $s \in \{0, ..., k\}$ is the unique stationary distribution of the Markov chain.

[10] Because there are two sides, we count this event twice.

[11] That the transition matrix is doubly stochastic can be seen from Equation 4, where the numbers pre-multiplying π_s on the right-hand side of the equation are the transition probabilities.

The Value Function Next, we provide an explicit characterization of the value function $\mathrm{W}(k)$. We then use this characterization to describe some of the properties of the function W.

Lemma 1. *If the planner's queue size is k, an arriving agent's steady-state payoff is*

$$\mathrm{W}(k) = \frac{H_c}{2k+1}\left[\left(k + x\frac{1-x^k}{1-x}\right) + l_c\right] - \frac{c}{r}. \tag{6}$$

Proof. We argue an arriving agent's expected payoff in steady state is

$$H\frac{k}{2k+1} + L\frac{1}{2k+1} + \sum_{i=0}^{k-1}\frac{u(i+1)}{2k+1}. \tag{7}$$

Equation 5 implies the probability at least one agent on the other side of the market is available to match is $\sum_{s=1}^{k}\pi_s = k/(2k+1)$. In this case, the arriving agent is immediately matched and receives a payoff of H, which explains the first term of Equation 7. With probability $\pi_k = 1/(2k+1)$, the arriving agent faces a queue of length k on her side of the market. Then, she leaves immediately and gets L, which explains the second term of Equation 7. Finally, the length of the queue on the side of the arriving agent is $i \in \{0, ..., k-1\}$ with probability $1/(2k+1)$, in which case, her expected payoff is $u(i+1)$ by Equation 2. Summing over $i \in \{0, ..., k-1\}$ yields the last term in Equation 7. Finally, we note Equation 7 yields Equation 6 after some algebraic manipulations. □

To determine the solution to the planner's problem, we analyze the planner's marginal value from having an additional agent join the queue, that is, $\mathrm{M}(k) \equiv \mathrm{W}(k+1) - \mathrm{W}(k)$. Lemma 1 implies the marginal value $\mathrm{M}(k)$ can be expressed as

$$\mathrm{M}(k) = H_c\frac{1 + x^{k+1} - 2l_c + 2x\sum_{i=0}^{k-1}(x^k - x^i)}{(2k+1)(2k+3)}. \tag{8}$$

Equation 8 illustrates the benefits and costs of increasing the queue size from k to $k+1$. On the one hand, increasing the queue size allows the planner to form an additional match by keeping an arriving agent when k agents are already queueing, when he would have otherwise sent this agent away. On the other hand, a larger queue size implies a randomly arriving agent faces a higher probability than before of having to queue in order to be served. Because waiting is costly, this shows up as a cost in Equation 8 (this is the term in the summation).

Lemma 2 is a key step in our analysis: the planner's marginal value from having an additional agent join the queue is decreasing whenever it is positive, and can cross 0 at most once. We use this result below to characterize the socially optimal queue size and, in Section 4, to compare the equilibrium and socially optimal queue sizes:

Lemma 2. *The following hold:*

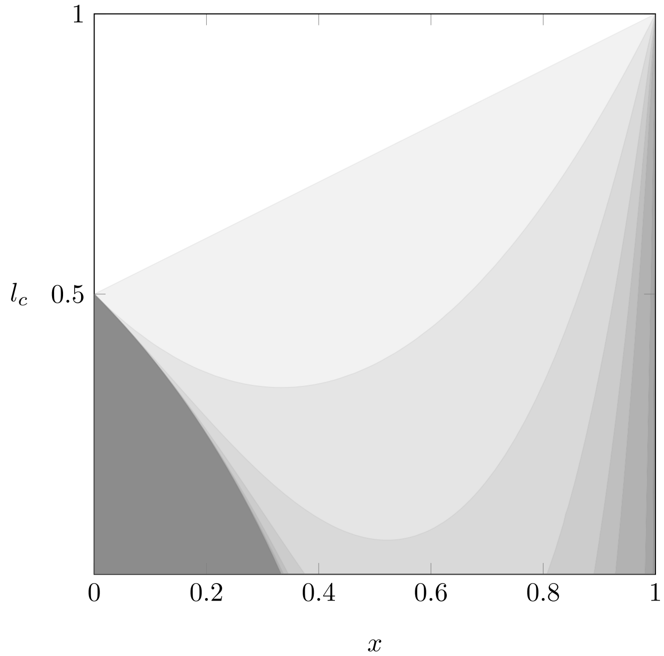

Figure 3: Socially optimal queue size, K^*.

(a) If $\mathrm{M}(k) < 0$*, then for all* $k' > k$, $\mathrm{M}(k') < 0$ *and*

(b) If $\mathrm{M}(k) \geq 0$*, then* $\mathrm{M}(k+1) \leq \mathrm{M}(k)$.

The proof of Lemma 2 and of other results in this section can be found in Appendix A.1.

Part (a) of Lemma 2 implies the socially optimal queue size, K^*, can be characterized as the threshold value at which the sign of the function M switches from positive to negative, that is,

$$\mathrm{M}(K^* - 1) > 0 \text{ and } \mathrm{M}(K^*) \leq 0. \tag{9}$$

Figure 3 displays the values of K^* as a function of (x, l_c), where the white area corresponds to $K^* = 0$, and darker shades of gray correspond to higher values of K^*.

Next, we establish a comparative static result regarding the socially optimal queue size. By Equation 8, the numerator of $\mathrm{M}(k)$ is decreasing in l_c. Hence, if $\mathrm{M}\,(k) \geq 0$ for a given l_c, this inequality continues to hold for any $l' \leq l_c$. By Equation 9, this implies the following:

Remark 3. *The socially optimal queue size,* K^**, is decreasing in* l_c.

Remark 3 formalizes the intuition that if agents have better outside options, the planner is less willing to have them join the queue.

Figure 3 shows K^* is not necessarily monotonic in x. This finding is in contrast to the observation in Remark 1 for k^*. To explain this difference, recall from Equation 9 that the change in the optimal queue size in response to an increase in x depends on the effect on $\mathrm{M}(k)$. By Equation 8, up to a factor H_c, we can express $\mathrm{M}(k)$ as

$$\frac{x^{k+1}}{2k+3} - 2\left[\sum_{i=0}^{k-1} \frac{x^{i+1}}{(2k+3)(2k+1)}\right] + \left[\frac{k}{2k+3} - \frac{k}{2k+1}\right] + \left[\frac{l_c+1}{2k+3} - \frac{l_c}{2k+1}\right].$$

An increase in x has the following two countervailing effects on the expression above. The first term, corresponding to the payoff of the agent who is last in line if the queue size is $k+1$, is increasing in x. The second term, corresponding to the change in payoffs of those agents who arrive in states $0, \ldots, k$ and do not match immediately, is also increasing in x. However, this term has a negative sign because the probability of each of these states goes down from $1/(2k+1)$ to $1/(2k+3)$. The final effect on K^* of an increase in x depends on the relative impact of x on the first and second terms. Figure 3 shows that, depending on the parameter values, either of these effects can dominate the other. The next remark shows that the net effect largely depends on l_c:

Remark 4. *The following hold:*

1. *If $l_c \geq 1/2$, K^* is non-decreasing in x.*
2. *If $l_c < 1/2$, K^* is U-shaped in x, that is, it is first decreasing and then increasing in x.*

To understand the statement in Remark 4, first understanding why $l_c = 1/2$ is the relevant cutoff is useful. Note that from each pair of arriving agents, one on each side of the economy, the planner can always guarantee a payoff of $2L$ by forcing both of them to take their outside option upon arrival. However, the planner can also guarantee a payoff of at least $H - c/r$ by forcing the first-to-arrive member of the pair to wait and match with the second member of the pair, who obtains a payoff of H. Note this scheme may entail the first agent waiting for an arbitrarily long period of time before being matched, making her payoff close to $-c/r$. Whenever

$$2L < H - \frac{c}{r} \Leftrightarrow 2L_c < H_c \Leftrightarrow l_c < \frac{1}{2},$$

x low enough exists that makes such a scheme profitable. However, when $l_c \geq 1/2$, the agent who waits should not wait for an arbitrarily long period of time; matching agents with their outside option upon arrival dominates the above scheme.[12]

The above discussion implies that, when $l_c \geq 1/2$, the planner chooses to keep a finite (possibly empty) queue.[13] Remark 4 shows that in this case, the socially optimal queue size is non-decreasing in x, and hence, the first of the two aforementioned effects dominates: the benefit from increased matching opportunities is greater than the increase in waiting costs. The reason

[12] This discussion highlights how the planner accounts for both sides of the market in his calculations (he compares $(1+x^k)H_c$ to $2L_c$), while the agents only compare their own benefit (they compare H_c or $x^k H_c$ to L_c).

[13] See Proposition 3.

is that for a fixed queue size, the planner always benefits from the increased arrivals and can keep the waiting costs at bay by not increasing the queue size. When $l_c < 1/2$, the planner may choose to keep an infinite-length queue when x is small enough. Then, as x increases, the increase in waiting costs generated by this policy is of first order, and thus, the socially optimal queue length initially diminishes. However, as x increases, the increase in matching opportunities eventually becomes first order in the planner's welfare calculations, and the socially optimal queue size increases in x.

4 Main results

In this section, we characterize the comparison between the equilibrium and socially optimal queue sizes in terms of the model parameters. As the analysis in Sections 2.1 and 3 shows, these queue sizes are determined by two parameters: the normalized outside option, l_c, which measures the value of the outside option relative to the value of matching in the queue, and the discounted arrival rate, x, which measures the expected discounted time until the next arrival.

To compare the equilibrium queue size, k^*, with the socially optimal one, K^*, we rely on Lemma 2. Lemma 2 implies that if $\mathrm{M}(k-1) \geq 0$ for $k \in \mathbb{N}$ then $K^* \geq k$. The reason is that, by part (b) of Lemma 2, the value of the social welfare function is increasing on $[0, k]$. On the flip side, if $\mathrm{M}(k) \leq 0$, part (a) of Lemma 2 implies the function W is decreasing on $[k, \infty)$, and hence, $K^* \leq k$. The results in this section rely on the above argument with $k = k^*$; that is, we evaluate the sign of the function M at the equilibrium queue size k^*. Indeed, we argue that evaluating the sign of M at $(\log l_c / \log x) - 1$ is sufficient. To see this, note the following. First, part (a) of Lemma 2 implies

$$\mathrm{M}\left(\frac{\log l_c}{\log x} - 1\right) \leq 0 \Rightarrow \mathrm{M}(k^*) \leq 0 \Rightarrow K^* \leq k^*,$$

as $\log l_c / \log x - 1 \leq k^*$. Second, part (b) of Lemma 2 implies

$$\mathrm{M}\left(\frac{\log l_c}{\log x} - 1\right) \geq 0 \Rightarrow \mathrm{M}(k^* - 1) \geq 0 \Rightarrow K^* \geq k^*,$$

as $\log l_c / \log x - 1 \geq k^* - 1$.

It follows that whether K^* is above or below k^* is determined by the sign of $\mathrm{M}(k-1)$ at $k = (\log l_c / \log x)$. Proposition 1 characterizes when $K^* \geq k^*$ as a function of (x, l_c). To introduce the result, define

$$\hat{l}(x) = \inf\left\{l \in [0,1] : \mathrm{M}\left(\frac{\log l}{\log x} - 1\right) \geq 0\right\}. \tag{10}$$

For each value of x, Equation 10 defines the smallest value of the normalized outside option

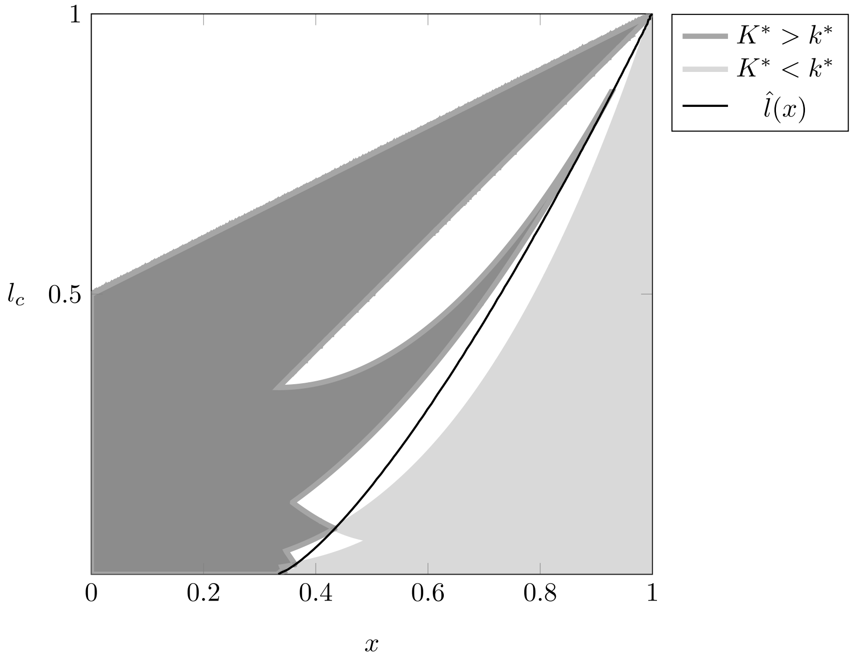

Figure 4: $K^* > k^*$ (dark gray), $K^* < k^*$ (light gray), $K^* = k^*$ (white), and the cutoff $\hat{l}(x)$ (black). The figure is based on numerical computations.

such that the numerator evaluated at $(\log l_c / \log x) - 1$ is non-negative. Inline with the previous discussion, $\hat{l}(x)$ is the smallest value of the outside option such that the socially optimal queue size dominates the equilibrium one.

We have the following:

Proposition 1 (Socially optimal vs. equilibrium queue sizes)**.** *For all $x \in [0, 1]$, $K^* \geq k^*$ if and only if $l_c \geq \hat{l}(x)$.*

Furthermore, the cutoff $\hat{l}(x)$ satisfies the following:

(a) For all $x \in [0, 1/3]$, $\hat{l}(x) = 0$,

(b) For all $x \in [0, 1]$, $\hat{l}(x) \leq x^2$, and

(c) For $x \geq 1/2$, $\hat{l}(x) \geq x^3$.

The first part of Proposition 1 states that the comparison between l_c and $\hat{l}(x)$ characterizes whether socially optimal queue sizes are larger than equilibrium ones. In particular, when the (normalized) outside option l_c is above the threshold $\hat{l}(x)$, the equilibrium queue is shorter than the socially optimal one. Remarks 1 and 3 imply both the equilibrium and socially optimal queue sizes are decreasing in l_c. Thus, the first part of Proposition 1 implies that as l_c increases, in a sense, the equilibrium queue size falls faster than the socially optimal one. To understand the underlying forces, consider the expression for the numerator of $\mathrm{M}(k)$ evaluated at $\log l_c/\log x - 1$:

$$\left[1 + x^{k+1} - 2l_c + 2x\sum_{i=0}^{k-1}(x^k - x^i)\right]|_{k=\frac{\log l_c}{\log x}-1} = 1 - l_c + 2\left[\left(\frac{\log l_c}{\log x} - 1\right) l_c + \frac{l_c}{1-x} - \frac{x}{1-x}\right].$$

The first part corresponds to the value of adding an additional agent to the queue $1+x^{k+1}-2l_c = 1-l_c$ as $x^{k^*+1} \approx l_c$. Because the higher the value of the outside option the higher the opportunity cost of having agents join the queue, the first part is always decreasing in l_c, yet it is positive as the planner internalizes the value to both sides of adding an additional agent to the queue. The term in square brackets corresponds to the change in waiting costs when increasing the queue size by 1 *starting from* the equilibrium queue size k^*. As l_c grows, this term becomes positive: This is intuitive because as l_c grows, the equilibrium queue size becomes small, and hence, the increase in costs of increasing the queue size by 1 is also small. Thus, when contemplating whether to add an additional agent to the queue, the planner internalizes the value of both sides, *and* as l_c grows, this additional agent is joining an already short queue. For this reason, as l_c grows, the planner becomes more patient than the agents.

The second part of Proposition 1 describes properties of the cutoff $\hat{l}$ that can be used to elucidate further comparisons between K^* and k^* (see also Proposition 2 below). Part (a) shows that when $x \leq 1/3$, the planner's queue size is at least as large as the equilibrium one for all values of the outside option l_c. Similarly, part (b) implies that when the equilibrium queue size is at most 1 (i.e., $l_c \geq x^2$), then $K^* \geq k^*$. Because x is increasing in λ and decreasing in r, parts (a) and (b) of Proposition 1 imply that if the arrival rate, λ, is small or the discount rate, r, is high, then equilibrium queues are too short relative to the efficient ones. In this case, waiting to be matched is too costly for the agents, and thus, they leave the market too early, ignoring that by doing so they reduce the matching opportunities encountered by future agents who arrive on the opposite side. By contrast, part (c) of Proposition 1 implies that when the arrival rate is high (i.e., $x \geq 1/2$) and equilibrium queue sizes are at least 3 (i.e., $l_c \leq x^3$), equilibrium queues are inefficiently long; that is, $K^* \leq k^*$.

Figure 4 illustrates in dark gray the parameter values (x, l_c) for which the socially optimal queue size is strictly above k^*, and in light gray, those for which it is strictly below k^*. Based on computations, Figure 4 illustrates that $\hat{l}$ is increasing in x. That is, fixing a value of l_c and increasing x, the planner starts off being more patient than the agents, that is, $K^* \geq k^*$, and then becomes more impatient, that is, $K^* < k^*$. To explain this observation, we argue that if x is small, the planner benefits from increasing the queue size at k^*, that is, $\mathrm{M}(k^*) \geq 0$. Similarly, if x is large, welfare is increased if the queue size is decreased at k^*, that is, $\mathrm{M}(k^*) \leq 0$. To this end, observe that if x is small, k^* is also small and the last term in the numerator of $\mathrm{M}(k^*)$ in Equation 8 is close to 0. Because $x^{k^*+1} \approx l_c$ (by Equation 3), the numerator can be approximated by $1 - l_c$, which is indeed positive, and hence, $\mathrm{M}(k^*) \geq 0$. By contrast, if x is close to 1, k^* is large and the last term in the numerator becomes a small negative number. In fact, this term is decreasing in x and can be arbitrarily small. Again, because $x^{k^*+1} \approx l_c$ (by Equation 3), the remaining terms in this numerator are around $1 - l_c$, which is dominated by the last term if x is big, showing $\mathrm{M}(k^*) \leq 0$. Indeed, Claim A.1 in Appendix A.2 shows that, for each l_c, $\mathrm{M}(k^* - 1) < 0$ for large enough x.

Proposition 2 refines Proposition 1 by providing conditions on the parameters (x, l_c) under which the socially optimal queue size K^* is strictly above or below k^*:

Proposition 2 (Socially optimal vs equilibrium queue sizes)**.** *The following hold:*

(a) If $x < l_c < (1+x)/2$, then $K^* > 0 = k^*$.

(b) If $l_c \leq x < 1/3$, then $K^* > 1 = k^*$.

(c) If $l_c < x^3$, and $x > 1/2$, then $k^* > K^* > 0$.

Proposition 3 below identifies parameter values for which the social planner always asks the agents to join the queue irrespective of its length, that is, $K^* = \infty$.

Proposition 3. *If $1 - 2l_c - 2x/(1-x) > 0$, then $K^* = \infty$.*

That the social planner finds it optimal to form arbitrarily long queues if l_c and x are small might appear surprising at first glance. Let us explain this observation. To do so, rewriting the condition in Proposition 3 as follows is useful:

$$H\left(1 - 2\frac{L}{H} - 2\frac{x}{1-x}\right) > \frac{c}{r}\left(\frac{1+x}{1-x}\right). \tag{11}$$

Assume first that agents are very impatient, that is, $r \approx \infty$, so that both x and c/r are approximately 0. In this case, whenever an agent has to wait, she receives a payoff close to zero. If the planner's queue size is K, then, by Equation 6, a newly arriving agent's payoff is roughly $(HK + L)/(2K+1)$. The reason is that the probability of having an agent on the other side of the market is $K/(2K+1)$, in which case, the agent is matched immediately and gets a payoff of H. With probability $1/(2K+1)$, a K-long queue on the agent's own side already exists, in which case, she takes her outside option and receives L. In any other case, the agent joins a queue and, because she is very impatient, receives a payoff close to zero. The expression $(HK + L)/(2K+1)$ (i) is strictly increasing in K if and only if $2L < H$, and (ii) converges to $H/2$ as K goes to infinity. Thus, by forming arbitrarily long queues, the planner maximizes the probability that a newly arriving agent does not have to wait. This probability cannot exceed $1/2$, so a newly arriving agent's payoff cannot be larger than $H/2$. Hence, the planner can find it optimal to set $K^* = \infty$ only if $H/2 > L$. Note that when $r \approx \infty$ the requirement $L < H/2$ is exactly the hypothesis of Proposition 3. Equation 11 shows that when the discount rate is finite, so that the planner cares about the agents' flow costs, the above scheme is socially optimal as long as it covers agents' costs of waiting indefinitely.

We emphasize that the possibility of unbounded queues is a consequence of the two-sidedness of the market we analyze. For illustration, consider a planner who only takes into account the welfare of agents on one side and can only determine the queue size for that side. The planner's problem then becomes similar to the one arising in a *server model*; see, for example, Naor (1969), where customers join a queue and are served as service opportunities arrive randomly. Then, the optimal queue size is always bounded; in fact, it is smaller than the largest equilibrium queue size. The reason is that, when determining one side's queue size, the planner ignores the positive externality an additional agent joining the queue imposes on the other side of the market. Instead, the planner does take into account the negative externality an additional agent joining the queue imposes on their same side. By contrast, unbounded queues can be socially optimal in our two-sided market because a queueing agent contributes to welfare not only through her payoff upon being matched, but also through the increase in speed at which agents are matched on the other side of the market.

Whereas Proposition 1 provides a full characterization of when the socially optimal queue size is larger or smaller than the equilibrium one, the characterization of when the comparison between K^* and k^* is strict in Propositions 2 and 3 is incomplete. Figures 5 and 6 below display our computational results regarding the comparison of k^* and K^*, together with the regions for which we have analytical results. Figure 5 depicts the regions identified by Propositions 2 and 3 where $K^* > k^*$. In both panels of Figure 5, the dark gray shaded area in both corresponds to $K^* \geq k^*$. The hatched black area in Figure 5a depicts the parameter values identified in parts (a) and (b) in Proposition 2. Similarly, the hatched black area in Figure 5b depicts the parameter values identified in Proposition 3, for which $K^* = \infty$. Instead, Figure 6 depicts the region identified by Proposition 2 where $K^* < k^*$: the shaded area in light gray corresponds to $K^* < k^*$ based on numerical computations, whereas the hatched area corresponds to part (c) in Proposition 2.

4.1 Equilibrium and socially optimal queue sizes under pure flow costs

In this section, we specialize our model to the canonical case of pure flow costs. That is, the agents experience only flow costs while they wait—$c > 0$—and do not discount the future—$r \approx 0$. The results herein allow us to compare our findings with those in the literature, especially those in Baccara et al. (2020).

In what follows, we characterize the equilibrium and socially optimal queue sizes, relying on the analysis in Section 2 and 3.

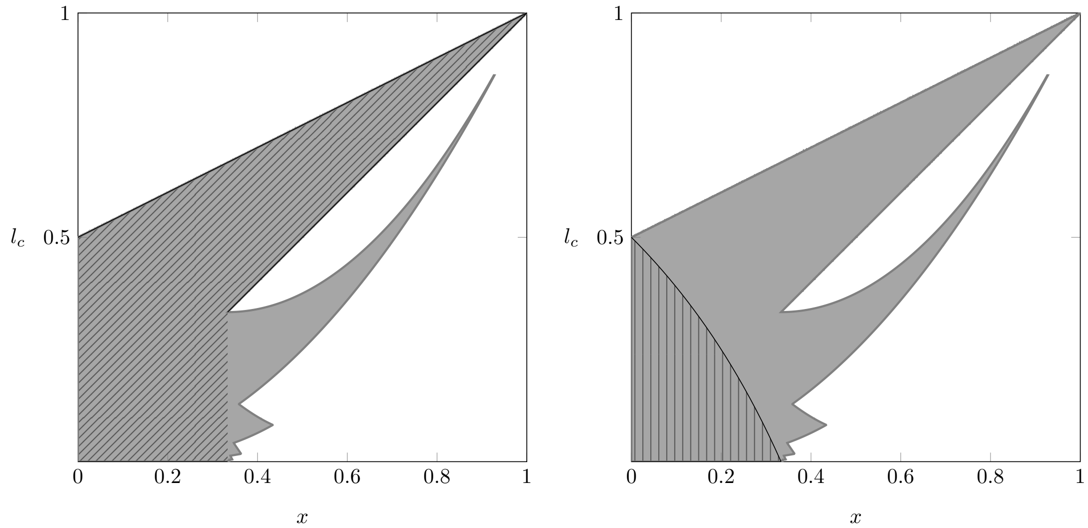

Figure (a) Proposition 2, (a) and (b) (hatched)

Figure (b) Proposition 3 (hatched)

Figure 5: $K^* > k^*$ (dark gray shaded). The dark gray shaded area is obtained via numerical computations and the hatched areas depict the conditions in Propositions 2 and 3.

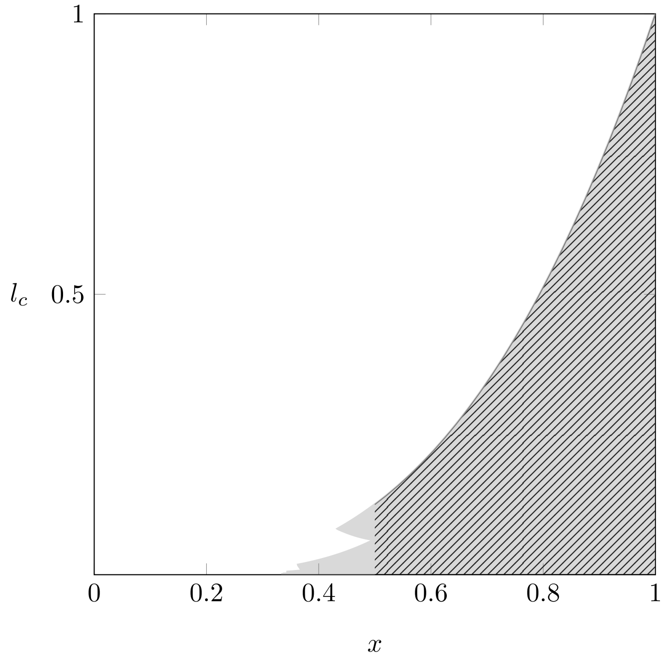

Figure 6: $K^* < k^*$(light gray shaded); Proposition 2, (c) (gray hatched). The light gray shaded area is obtained via numerical computations.

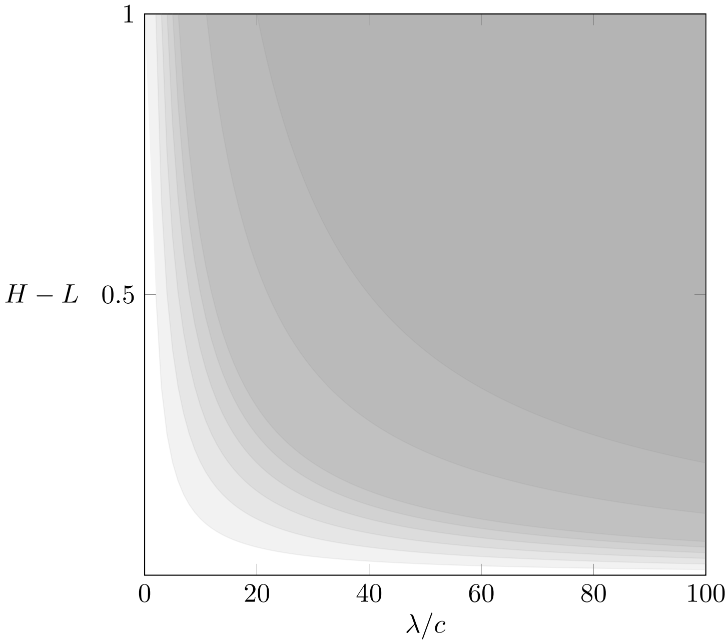

Figure 7: Equilibrium queue size, $\overline{k}^*$.

Equilibrium queue size Setting $r = 0$ in Equation 1 delivers that the payoff $\overline{u}(k)$ of an agent who joins a queue of length $k-1 \geq 0$ satisfies that

$$\overline{u}(k) = \overline{u}(k-1) - \frac{c}{\lambda},$$

where again $\overline{u}(0) = H$.

Solving for $\overline{u}$ by induction yields

$$\overline{u}(k) = H - k\frac{c}{\lambda}.$$

Therefore, an agent is willing to join a queue of size $k-1$ if and only if

$$H - L \geq k\frac{c}{\lambda}.$$

Thus, the equilibrium queue size is determined by

$$\overline{k}^* = \left\lfloor \frac{(H-L)\lambda}{c} \right\rfloor.$$

Figure 7 depicts the equilibrium queue size $\overline{k}^*$ for $(\lambda/c, H-L) \in [0,100] \times [0,1]$, where the white area corresponds to $\overline{k}^* = 0$ and darker shades of gray correspond to higher values of $\overline{k}^*$.

Socially optimal queue size We denote the welfare and marginal welfare functions by $\overline{\mathrm{W}}$ and $\overline{\mathrm{M}}$, respectively. Similar arguments to those in the derivations of W and M imply

$$\overline{\mathrm{W}}(k) = \frac{1}{2k+1}\left[L + 2Hk - \frac{c}{\lambda}\frac{k(k+1)}{2}\right],$$

and

$$\overline{\mathrm{M}}(k) \equiv \overline{\mathrm{W}}(k+1) - \overline{\mathrm{W}}(k) = \frac{1}{(2k+1)(2k+3)}\left[2(H-L) - \frac{c}{\lambda}(k+1)^2\right].$$

The socially optimal queue size, $\overline{K}^*$, satisfies:

$$\overline{\mathrm{M}}(\overline{K}^* - 1) \geq 0 \text{ and } \overline{\mathrm{M}}(\overline{K^*}) \leq 0.$$

Thus, $\overline{K}^*$ satisfies

$$\sqrt{\frac{2(H-L)\lambda}{c}} - 1 \leq K^* \leq \sqrt{\frac{2(H-L)\lambda}{c}}.$$

Figure 8 below depicts the planner's cutoff for $(\lambda/c, H-L) \in [0,100] \times [0,1]$, where the white area corresponds to $\overline{K}^* = 0$ and darker shades of gray correspond to higher values of $\overline{K}^*$.

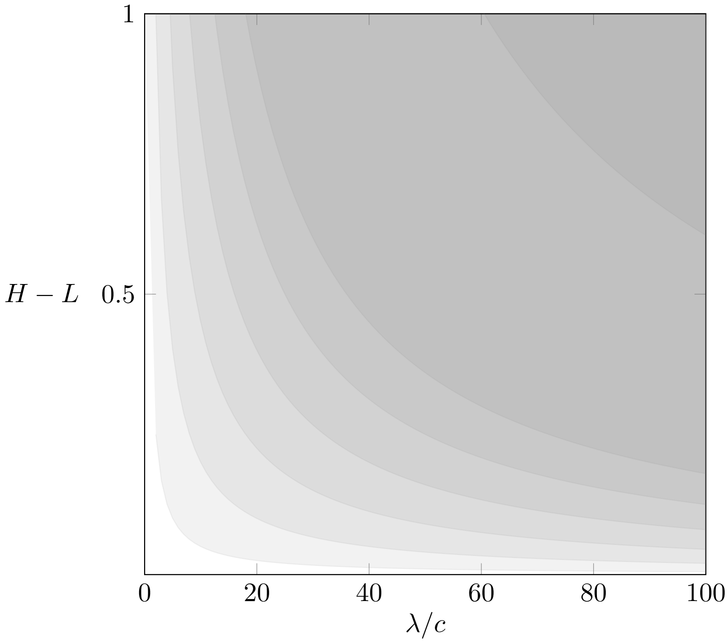

Figure 8: Socially optimal queue size, $\overline{K}^*$.

Proposition 4 is the main result of this section:

Proposition 4 (Socially optimal vs equilibrium queue size under pure flow costs)**.** *When agents incur only flow costs while they wait in the queue, the following holds:*

(i) If $\overline{k}^* \geq 1$, then $\overline{K}^* \leq \overline{k}^*$ and the inequality is strict whenever $\overline{k}^* \geq 3$.

(ii) Parameter values exist such that $\overline{K}^* = 1$ and $\overline{k}^* = 0$.

The proof is in Appendix A.3. To see why the result holds, ignore the integer constraints on $\overline{K}^*$ and $\overline{k}^*$. That is, take $\overline{K}^* = \lambda(H-L)/c$ and $\overline{K}^* = \sqrt{2\lambda(H-L)/c}$. Then, $\overline{k}^* = \left(\overline{K}^*\right)^2/2$ and, therefore,

$$\overline{k}^* - \overline{K}^* \geq 0 \Leftrightarrow \overline{K}^*(\overline{K}^* - 2) \geq 0 \Leftrightarrow \overline{K}^* \geq 2,$$

Thus, once the socially optimal queue entails at least two agents, the equilibrium queue length is always inefficiently high. Moreover, when the individual queue length is exactly 1, that is, when $\lambda(H-L)/c \in [1,2)$, $\overline{K}^* = 1$.[14] However, when the socially optimal queue has length 1, parameter values can be found such that every agent takes the outside option upon arrival.

Proposition 4 implies the planner generally keeps shorter queues than those that arise in equilibrium in the pure flow costs case. This finding is in sharp contrast to our results regarding discounting agents in Section 4. To explain this observation, note the model in Section 2 differs from the pure flow costs model in two ways, both of which suggest the planner will hold shorter queues in the case of flow costs:

First, in the pure flow costs model, the planner is reluctant to maintain long queues, because an agent's payoff goes to minus infinity as her waiting time converges to infinity. Consequently, the cost of forcing an additional agent to join the queue to create another match is prohibitively high if the queue is already long. In the case of discounting, an agent's payoff from waiting to be matched is at least $-c/r$ irrespective of her waiting time. Therefore, the cost of forcing an additional agent to join the queue is bounded, which, in turn, might result in arbitrarily long queues (see Proposition 3).

The second difference relates to how discounting and pure flow costs shape agents' decisions to join the queue. To illustrate, we compare the cases of pure discounting ($r > 0, c = 0$) and pure flow costs ($r = 0, c > 0$).[15] Let us define the net value of matching of an agent as the difference between the payoff of matching evaluated at the moment of joining the queue and the outside option, L. Discounting introduces a wedge between the net value of matching of an agent who finds a queue of size k on her side, $Hx^{k+1} - L$, and that of an agent who finds a queue of size k on the opposite side, $H - L$. Both these values are different from the net value of matching

[14] To see this, note $\lambda(H-L)/c \in [1,2)$ implies $\sqrt{2\lambda(H-L)/c} \in [\sqrt{2}, 2)$, and hence, $\overline{K}^* = 1$.

[15] Remark 2 implies the same argument holds when the agents discount the future and incur in flow costs while they wait. By appropriately replacing H and L by H_c and L_c in this paragraph's discussion, the same argument goes through.

of a randomly arriving agent when k agents are in the system, $(H + Hx^{k+1} - 2L)/2$. Note the latter corresponds to the first part of the numerator in $\mathrm{M}(k)$, which is the positive term in the marginal social value of increasing the queue from k to $k+1$. By contrast, in the pure flow costs case, the net value of matching is the same in both cases and coincides with that of a randomly arriving agent in the system, $2(H-L)/2 = H-L$, which is the same as the positive term in the numerator in $\overline{\mathrm{M}}(k)$. Thus, whereas in both models, the agents and the planner "disagree" on how to compare the negative effects generated by joining the queue (captured by the negative terms in $\mathrm{M}(k)$ and $\overline{\mathrm{M}}(k)$, respectively), in the case of pure flow costs, they "agree" on how to compare the positive effects generated by joining the queue, conditional on the agents joining the queue in equilibrium.

To compare our results with those in Baccara et al. (2020), let us briefly describe their model. Baccara et al. (2020) consider a discrete-time, dynamic matching environment with two sides. Unmatched agents on the market pay a flow cost of c. Each agent has a binary quality, either high or low. Low-quality agents in their model play the role of our outside option: all agents prefer to match with a high-quality agent and match with a low-quality agent only when the waiting time for a high-quality agent is high enough. A new agent arrives on each side of the market every period. Baccara et al. (2020) show equilibrium queues are *always* longer queues than the socially optimal ones. Proposition 4 confirms their result, except that in our model, equilibrium queues may be empty when having a queue of size 1 is socially optimal. In our model, if the equilibrium queue size is 0 and the planner forces an agent to form a queue of length 1, this agent generates a positive externality for the other side because the next agent arriving on the other side obtains a payoff of H instead of L. In addition, this agent does not increase the waiting time of agents arriving on her side, because they take their outside options.[16] Thus, the planner may set $\overline{K}^* = 1$, despite $\overline{k}^* = 0$. In the model of Baccara et al. (2020), making a high-quality agent wait to be matched with a high-quality agent on the other side, instead of matching her with the newly arrived low-quality agent immediately, has two effects. On the one hand, the waiting agent generates a future benefit for an arriving high-quality agent on the other side, who would have been matched with a low-quality agent otherwise. On the other hand, she generates a cost for the low-quality agent on the other side who is now forced to wait, but would have been matched immediately otherwise. Thus, even though this queueing agent does not increase waiting costs on her side, she does so on the other side of the market. This cost overturns the benefit of keeping a queue of length 1, whenever forming a queue of length 1 is not an equilibrium. Finally, notice that as long as the socially optimal queue is at least 2, the equilibrium in our model entails inefficiently long queues, as in Baccara et al. (2020). Now, when the planner is evaluating whether to add an agent to an existing queue of length k, he considers not only the benefit for agents on the other side, but also the increased waiting times for agents

[16] The planner compares $2(H-L)$ (benefit) to c/λ (cost). In contrast, an individual agent only compares her benefit from waiting, $H-L$, with her waiting cost, c/λ, failing to internalize the benefit she generates for the other side.

on the same side, which is a product of the longer queues.

5 Extensions: Asymmetric markets

The analysis so far presumes both sides are symmetric: they have equal preferences and arrival rates. In this section, we consider the implications for our analysis of the agents having different outside options (Section 5.1) and arrival rates (Section 5.2). In what follows, distinguishing between the two sides of the market is useful, and we label them A and B.

5.1 Asymmetric outside options

In this section, we assume the outside options on each side are different. Letting L_A denote the value of the outside option on side A and L_B that on side B, we assume without loss that $L_A < L_B$. As in the model in Section 2, equilibrium and socially optimal queue sizes depend on the ratios

$$l_c^A \equiv \frac{L_A + \frac{c}{r}}{H + \frac{c}{r}}, \ \ l_c^B \equiv \frac{L_B + \frac{c}{r}}{H + \frac{c}{r}}.$$

In fact, denote by $k^*(l_c)$ and $K^*(l_c)$ the equilibrium and socially optimal queue sizes derived in Sections 2.1 and 3, respectively. As we show below, these objects are enough to characterize the equilibrium and socially optimal queue sizes when outside options are different across sides.

Equilibrium queue sizes Agents' equilibrium queueing decisions are determined as in Section 2.1, except that the queues on each side may have different lengths. In particular,

$$k_A^* = \left\lfloor \frac{\log l_c^A}{\log x} \right\rfloor = k^*(l_c^A) \quad \text{and } k_B^* = \left\lfloor \frac{\log l_c^B}{\log x} \right\rfloor = k^*(l_c^B). \tag{12}$$

Because side B agents have better outside options, we naturally conclude that $k_B^* \leq k_A^*$.[17]

Socially optimal queue size(s) As in Section 3, we restrict the planner to choosing the queue size. Unlike in Section 3, we allow the planner to select different queue sizes k_A, k_B for each side.

Instead of modeling the planner as choosing two queues with sizes k_A and k_B, thinking of the planner as operating a single queue whose length can be any $k \in \{-k_A, \ldots, -1, 0, 1, \ldots, k_B\}$ is useful. In this single queue, negative lengths indicate the side A queue is non-empty and positive lengths indicate the side B queue is non-empty. The planner's thresholds, k_A and k_B, together with the arrival probability, λ, induce a distribution over the different queue lengths. As in

[17] Note $k_B^* = 0 < k_A^*$ is possible. In this case, agents on side B match only if they find an agent on side A already queueing when they arrive; otherwise, they leave unmatched.

Section 3, we can show that the ergodic distribution satisfies

$$\pi_s = \frac{1}{k_A + k_B + 1},$$

for each $s \in \{-k_A, \ldots, 0, \ldots, k_B\}$.

It follows that the welfare of a randomly arriving agent on each side of the market is given by

$$\hat{\mathrm{W}}_A(k_A, k_B) = \frac{H_c}{k_A + k_B + 1}\left[k_B + x\frac{1 - x^{k_A}}{1 - x} + l_c^A\right] - \frac{c}{r},$$
$$\hat{\mathrm{W}}_B(k_A, k_B) = \frac{H_c}{k_A + k_B + 1}\left[k_A + x\frac{1 - x^{k_B}}{1 - x} + l_c^B\right] - \frac{c}{r},$$

and hence, total welfare is given by

$$\begin{aligned}\hat{\mathrm{W}}(k_A, k_B) &= \frac{1}{2}\hat{\mathrm{W}}_A(k_A, k_B) + \frac{1}{2}\hat{\mathrm{W}}_B(k_A, k_B) \\ &= \frac{H_c/2}{k_A + k_B + 1}\left[k_A + k_B + x\frac{2 - x^{k_A} - x^{k_B}}{1 - x} + l_c^A + l_c^B\right] - \frac{c}{r},\end{aligned}$$

where the hat notation $\hat{\mathrm{W}}$ distinguishes the planner's value function from that in Section 2.

Proposition 5 characterizes the socially optimal queue sizes. Even though the planner can set different queue sizes for each side, we find doing so is not optimal:

Proposition 5 (Optimality of symmetric queues)**.** *The socially optimal queue sizes, K_A^*, K_B^*, satisfy*

$$K_A^* = K_B^* = K^*\left((l_c^A + l_c^B)/2\right).$$

The proof of this and of other results in this section are in Appendix A.4.

In words, the planner's optimal queue size corresponds to that of a market in which agents on both sides have symmetric outside options with value $l_c^P \equiv \left(l_c^A + l_c^B\right)/2$. To see why the result holds, ignore the integer constraints on k_A and k_B, and note the planner's value function satisfies the following:

$$\begin{aligned}\hat{\mathrm{W}}(k_A, k_B) &= \frac{H_c}{2\frac{k_A + k_B}{2} + 1}\left[\frac{k_A + k_B}{2} + x\frac{1 - \frac{x^{k_A} + x^{k_B}}{2}}{1 - x} + \frac{l_c^A + l_c^B}{2}\right] - \frac{c}{r} \qquad (13)\\ &\leq \frac{H_c}{2\frac{k_A + k_B}{2} + 1}\left[\frac{k_A + k_B}{2} + x\frac{1 - x^{\frac{k_A + k_B}{2}}}{1 - x} + \frac{l_c^A + l_c^B}{2}\right] - \frac{c}{r} = \mathrm{W}\left(\frac{k_A + k_B}{2}; l_c^P\right),\end{aligned}$$

where the inequality follows from x^k being convex in k and $\mathrm{W}(\cdot, l_c^P)$ is the planner's value function in the model in Section 3 when the outside option is l_c^P.

In words, the value of queue sizes (k_A, k_B) in the asymmetric outside-options market is bounded above by the value of queue size $(k_A + k_B)/2$ in the symmetric market with outside option l_c^P. Moreover, both values coincide whenever $k_A = k_B$. The convexity of x^k in k implies the planner dislikes setting unequal queue sizes, because doing so leads to greater variability in payoffs when the agents queue to be matched.

Proposition 5 together with the results in Section 4 imply the following:

Corollary 1 (Equilibrium vs. socially optimal queue sizes with asymmetric outside options)**.** *The following hold:*

1. *If $l_c^P \geq \hat{l}(x)$, the socially optimal queue size is longer than side B's equilibrium queue size, that is, $K^*(l_c^P) \geq k_B^*$,*
2. *If $l_c^P \leq \hat{l}(x)$, the socially optimal queue size is shorter than side A's equilibrium queue size, that is, $K^*(l_c^P) \leq k_A^*$,*
3. *If $1 - 2l_c^P - 2x/(1-x) > 0$, then $K^*(l_c^P) = \infty$.*

Because of the asymmetric outside options, it is possible that $k_B^* \leq K^*(l_c^P) \leq k_A^*$, so that the planner is neither less nor more patient than both sides of the market. The argument after Proposition 5 suggests, however, one would like to compare the planner's queue size $K^*(l_c^P)$ with the average of equilibrium queue sizes $(k_A^* + k_B^*)/2$ (cf. Equation 13). Whereas establishing this comparison in general is complicated because of the integer constraints, as we show next, in the pure flow costs case, the planner typically keeps queues shorter than $(k_A^* + k_B^*)/2$:

Proposition 6 (Average equilibrium queues are too long when $r = 0$)**.** *Suppose agents only incur in flow costs, that is, $r = 0$, and let $\overline{k}_A^*, \overline{k}_B^*$ denote the equilibrium queue sizes. Then, the following hold:*

1. *If $(\overline{k}_A^* + \overline{k}_B^*) \geq 2$, the socially optimal queue size is shorter than the average equilibrium queue size.*
2. *Instead, if $(\overline{k}_A^* + \overline{k}_B^*) \leq 1$, the socially optimal queue size may be longer than the average equilibrium queue size. In particular, whenever $0 = \overline{k}_B^* < 1 = \overline{k}_A^*$, the socially optimal queue length is 1.*

Proposition 6 is analogous to Proposition 4, except that now one must account for the potentially asymmetric queue length on each side of the market. Indeed, the main difference relative to Proposition 4 is that the planner may find it optimal to have both sides queue even if only side A agents find it optimal to queue in equilibrium. As we show in the appendix, the planner also dislikes unequal queue sizes in the pure flow costs case.

5.2 Asymmetric arrival rates

In this section, we assume the arrival rates on each side are different. Letting λ denote the arrival rate on side A and μ that on side B, we assume without loss that $\mu > \lambda$. In what follows, the ratio λ/μ is important and we denote it by ρ.

Equilibrium queue sizes To define the equilibrium queue sizes, let

$$x_A = \frac{\mu}{\mu + r}, x_B = \frac{\lambda}{\lambda + r},$$

denote the discounted arrival rates on each side. Note that because agents on side B arrive more frequently, it is as if agents on side A are more patient. In what follows, we assume $x_B < \rho$, or equivalently, that $\mu < \lambda + r$.

Similar to the analysis in Section 2.1, we conclude the equilibrium queue sizes for each side are given by

$$k_A^* = \left\lfloor \frac{\log l_c}{\log x_A} \right\rfloor \quad \text{and} \quad k_B^* = \left\lfloor \frac{\log l_c}{\log x_B} \right\rfloor . \tag{14}$$

Remark 1 implies $k_A^* \geq k_B^*$: because arrivals on side B are more frequent, waiting times on side A are shorter, and hence, agents on side A are more willing to join the queue on their side.

Planner's problem Like in Section 5.1, defining the planner's queue as having negative and positive lengths $k \in \{-k_A, \ldots, 0, \ldots, k_B\}$ is useful. Unlike in the previous sections, the ergodic distribution induced by the planner's queue-size policy (k_A, k_B) and the arrival rates is not uniform across states. Consider an interval of time of length Δ. The following three heuristic equations establish the relationship between the ergodic probabilities of different states:

$$\begin{aligned}
\pi_{-k_A} &= (1 - \mu\Delta)\pi_{-k_A} + \lambda\Delta\pi_{-k_A+1} \\
\pi_s &= (1 - (\mu + \lambda)\Delta)\pi_s + \mu\Delta\pi_{s-1} + \lambda\Delta\pi_{s+1} \ , \ s \in \{-k_A + 1, \ldots, 1, \ldots, k_B - 1\} \\
\pi_{k_B} &= (1 - \lambda\Delta)\pi_{k_B} + \mu\Delta\pi_{k_B-1}.
\end{aligned}$$

Working through these equations delivers the following system as a function of the probability of the state k_B, π_{k_B},

$$\pi_{k_B - s} = \rho^s \pi_{k_B}, \ \ s \in \{0, \ldots, k_B + k_A\}.$$

Because agents on side B arrive more frequently, the state in which the side B queue is full, k_B, has the highest probability. The probability of the remaining states is obtained by geometrically discounting that of state k_B by ρ, and the smaller the number of agents on side B in the market at that state, the larger the discounting.

Noting $\sum_{s=-k_A}^{k_B} \pi_s = 1$, we obtain that the unique ergodic distribution is given by

$$\pi_{k_B - s} = \rho^s \pi_{k_B} = \rho^s \frac{1-\rho}{1-\rho^{k_A+k_B+1}}, \quad s \in \{0, \ldots, k_B + k_A\}.$$

Note the ergodic distribution is only defined for $k_B < \infty$. This feature already highlights a difference from our previous analysis: with asymmetric arrival rates, we cannot have unbounded queues on both sides.

Planner's value function It follows that the welfare of a randomly arriving agent on sides A and B is given by

$$\begin{aligned} \hat{\mathrm{W}}_A(k_A, k_B) &= \pi_{k_B} H_c \left[\frac{1-\rho^{k_B}}{1-\rho} + \rho^{k_B} x_A \frac{1-(\rho x_A)^{k_A}}{1-\rho x_A} + \rho^{k_A+k_B} l_c \right] - \frac{c}{r}, \\ \hat{\mathrm{W}}_B(k_A, k_B) &= \pi_{k_B} H_c \left[\rho^{k_B+1} \frac{1-\rho^{k_A}}{1-\rho} + \rho^{k_B} x_B \frac{1-(x_B/\rho)^{k_B}}{1-x_B/\rho} + l_c \right] - \frac{c}{r}. \end{aligned} \tag{15}$$

The asymmetry in the arrival rates translates into asymmetry in the welfare of each of the sides even if $k_A = k_B$. For instance, because the system is more likely to be in state k_B than state $-k_A$, agents on side B are more likely to take their outside option than agents on side A. For this reason, the outside option term in $\hat{\mathrm{W}}_A$ is pre-multiplied by $\rho^{k_A+k_B}$. Similarly, agents on side A are more likely to match upon arrival than agents on side B, which explains why the first term in $\hat{\mathrm{W}}_B$ is pre-multiplied by ρ^{k_B+1}, whereas the analogous term in $\hat{\mathrm{W}}_A$ is not.

Equation 15 implies the steady-state payoff of a randomly arriving agent is given by

$$\begin{aligned} \hat{\mathrm{W}}(k_A, k_B) &= \frac{1}{1+\rho} W_B(k_A, k_B) + \frac{\rho}{1+\rho} W_A(k_A, k_B) \\ &= \frac{H_c \pi_{k_B}}{1+\rho} \left[l_c(1+\rho^{k_A+k_B+1}) + \frac{\rho(1-\rho^{k_A+k_B})}{1-\rho} + \rho^{k_B} \left(x_B \frac{1-(x_B/\rho)^{k_B}}{1-(x_B/\rho)} + \rho x_A \frac{1-(\rho x_A)^{k_A}}{1-\rho x_A} \right) \right] - \frac{c}{r}. \end{aligned} \tag{16}$$

The payoff of a randomly arriving agent is described by three terms: the first is the welfare from taking the outside option; the second is the welfare of matching upon arrival; and the third is the welfare of matching after joining a queue.

In Appendix A.5, we show the marginal benefits of adding an agent on either side of the queue,

$$\hat{\mathrm{M}}_A(k_A, k_B) \equiv \hat{\mathrm{W}}(k_A+1, k_B) - \hat{\mathrm{W}}(k_A, k_B) \text{ and } \hat{\mathrm{M}}_B(k_A, k_B) \equiv \hat{\mathrm{W}}(k_A, k_B+1) - \hat{\mathrm{W}}(k_A, k_B),$$

satisfy properties analogous to those established in Lemma 2. In particular, for each side B queue size, k_B, an optimal side A queue size $K_A^*(k_B)$ exists, which is determined by the condition

$$\hat{\mathrm{M}}_A(K_A^*(k_B) - 1, k_B) > 0 \text{ and } \hat{\mathrm{M}}_A(K_A^*(k_B), k_B) \leq 0.$$

Similarly, for each side A queue size, k_A, an optimal side B queue size $K_B^*(k_A)$ exists. This observation is an important step toward characterizing the socially optimal queue sizes, $(\hat{K}_A^*, \hat{K}_B^*)$, because they satisfy that $K_A^*(\hat{K}_B^*) = \hat{K}_A^*$ and $K_B^*(\hat{K}_A^*) = \hat{K}_B^*$.

Admission control in double-ended queues is a largely unexplored topic, and characterizing the socially optimal queue sizes is outside this paper's scope.[18] Instead, we use the above observations to illustrate how some of our results in the previous sections change once we consider asymmetric arrival rates.[19]

Proposition 7. *If*

$$1 - 2l_c - \frac{\rho x_A}{1 - \rho x_A} - x_B \frac{\frac{\ln x_B}{\ln \rho} - 1}{1 - x_B/\rho} > 0,$$

the socially optimal queue sizes are $(\hat{K}_A^*, \hat{K}_B^*) = (+\infty, 0)$.

The proof of Proposition 7 is in Appendix A.5. Similar to Proposition 3, Proposition 7 is obtained by analyzing the limiting behavior of the side A marginal benefit $\hat{\mathrm{M}}_A$. In fact, the condition in Proposition 7 guarantees that for each k_B, the planner always prefers to set $K_A^*(k_B) = +\infty$. With this condition at hand, we obtain the candidate for the optimal queue size on side B from side B's marginal benefit function $\hat{\mathrm{M}}_B$. We then show analytically that when $\mu \geq 2\lambda$, $\hat{K}_B^* = 0$, that is, when side B arrives much more frequently than side A, the planner does not allow a queue to form on side B. Finally, we verify numerically that when $\mu < 2\lambda$, the condition in Proposition 7 implies the optimal queue size on side B is 0.

We note the contrast between this result and those in Section 5.1. First, the planner is willing to choose a maximally asymmetric queue under the conditions in Proposition 7, whereas Proposition 5 implies doing so would not be optimal under asymmetric outside options.[20] Second, in contrast to Corollary 1, Proposition 7 provides a condition under which the planner is more patient than the more patient side in the market—side A—and more impatient than the more impatient one—side B. Indeed, by having agents on side A join the queue with probability 1, the planner maximizes the probability that an arriving agent on side B is matched upon arrival. Instead, because agents on side B arrive more often than those on A, whenever an agent on side B arrives to an empty queue on side A, the planner prefers them to take their outside option.

[18] Recent studies of admission control use heavy traffic approximations (Liu and Weerasinghe, 2021) or assume queue sizes are exogenously bounded (Su and Li, 2023), commenting on the difficulty of the analysis. Instead, acknowledging the difficulties, Liu et al. (2022) focus on developing algorithms to study admission control. None of these papers deal with strategic behavior.

[19] In their extension to asymmetric arrival rates, Baccara et al. (2020) also characterize the optimal queue size for the fast arriving side when having an unbounded queue on the slow arriving side is optimal.

[20] An incomplete intuition for the difference across the models follows from noting that the payoffs from waiting to be matched, $\sum_{i=0}^{k-1} x^i$, are concave in k but convex in x. Thus, when the discounted arrival rate is the same across sides as in Section 5.1, the planner prefers symmetric queues, but this force is not present once the discounted arrival rates differ.

References

Adan, I. and G. Weiss (2012): "Exact FCFS matching rates for two infinite multitype sequences," *Operations research*, 60, 475–489.

Afèche, P., A. Diamant, and J. Milner (2014): "Double-sided batch queues with abandonment: Modeling crossing networks," *Operations Research*, 62, 1179–1201.

Agarwal, N., I. Ashlagi, M. A. Rees, P. Somaini, and D. Waldinger (2021): "Equilibrium allocations under alternative waitlist designs: Evidence from deceased donor kidneys," *Econometrica*, 89, 37–76.

Akbarpour, M., S. Li, and S. O. Gharan (2020): "Thickness and information in dynamic matching markets," *Journal of Political Economy*, 128, 783–815.

Anderson, R., I. Ashlagi, D. Gamarnik, and Y. Kanoria (2015): "A dynamic model of barter exchange," in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 1925–1933.

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

Atakan, A. E. (2006): "Assortative matching with explicit search costs," *Econometrica*, 74, 667–680.

Baccara, M., S. Lee, and L. Yariv (2020): "Optimal dynamic matching," *Theoretical Economics*, 15, 1221–1278.

Bell, C. E. (1971): "Characterization and computation of optimal policies for operating an M/G/1 queuing system with removable server," *Operations Research*, 19, 208–218.

Boxma, O. J., I. David, D. Perry, and W. Stadje (2011): "A new look at organ transplantation models and double matching queues," *Probability in the Engineering and Informational Sciences*, 25, 135–155.

Che, Y.-K. and O. Tercieux (2021): "Optimal queue design," in *Proceedings of the 22nd ACM Conference on Economics and Computation*, 312–313.

Das, S., J. P. Dickerson, Z. Li, and T. Sandholm (2015): "Competing dynamic matching markets," in *Proceedings of the Conference on Auctions, Market Mechanisms and Their Applications (AMMA)*, vol. 112, 245.

Diamant, A. and O. Baron (2019): "Double-sided matching queues: Priority and impatient customers," *Operations Research Letters*, 47, 219–224.

Dobbie, J. M. (1961): "A doubled-ended queuing problem of Kendall," *Operations Research*, 9, 755–757.

Doval, L. (Forthcoming): "Dynamic Matching," in *Handbook of the Economics of Matching*, ed. by Y.-K. Che, P.-A. Chiappori, and B. Salanie, Elsevier.

Elster, J. (1992): *Local justice: How institutions allocate scarce goods and necessary burdens*, Russell Sage Foundation.

Forum, B. (2019): "Role lock forces more high elo players to smurf," https://us.forums.blizzard.com/en/overwatch/t/role-lock-forces-more-high-elo-players-to-smurf/417771, accessed: September 11, 2024.

Gurvich, I. and A. Ward (2015): "On the dynamic control of matching queues," *Stochastic Systems*, 4, 479–523.

Hassin, R. (1985): "On the optimality of first come last served queues," *Econometrica*, 53, 201–02.

Huitfeldt, I., V. Marone, and D. C. Waldinger (2024): "Designing Dynamic Reassignment Mechanisms: Evidence from GP Allocation," Tech. rep., National Bureau of Economic Research.

Kendall, D. G. (1951): "Some problems in the theory of queues," *Journal of the Royal Statistical Society: Series B (Methodological)*, 13, 151–173.

Larson, R. C. (1987): "OR forum—perspectives on queues: Social justice and the psychology of queueing," *Operations research*, 35, 895–905.

Leshno, J. D. (2022): "Dynamic matching in overloaded waiting lists," *American Economic Review*, 112, 3876–3910.

Lindsay, C. M. and B. Feigenbaum (1984): "Rationing by waiting lists," *The American economic review*, 74, 404–417.

Liu, H.-L., Q.-L. Li, Y.-X. Chang, and C. Zhang (2022): "Double-ended queues with non-Poisson inputs and their effective algorithms," *Computers & Operations Research*, 144, 105793.

Liu, X. and A. Weerasinghe (2021): "Admission control for double-ended queues," *arXiv preprint arXiv:2101.06893*.

Naor, P. (1969): "The regulation of queue size by levying tolls," *Econometrica: journal of the Econometric Society*, 15–24.

Nguyen, H. Q. and T. Phung-Duc (2022a): "Queueing Analysis and Nash Equilibria in an Unobservable Taxi-passenger System with Two Types of Passenger." in *ICORES*, 48–55.

——— (2022b): "A Two-Population Game in Observable Double-Ended Queueing Systems," *Operations Research Letters*, 50, 407–414.

Nikzad, A. and P. Strack (2024): "Equity and Efficiency in Dynamic Matching: Extreme Waitlist Policies," *Management Science*, 70, 5187–5207.

Prendergast, C. and L. Stole (2001): "Barter, liquidity and market segmentation," *Available at SSRN 287848*.

Ross, S. M. (2010): *Introduction to probability models*, Academic press.

Schummer, J. (2021): "Influencing waiting lists," *Journal of Economic Theory*, 195, 105263.

Shi, Y. and Z. Lian (2016): "Optimization and strategic behavior in a passenger–taxi service system," *European Journal of Operational Research*, 249, 1024–1032.

Shimer, R. and L. Smith (2000): "Assortative matching and search," *Econometrica*, 68, 343–369.

——— (2001): "Matching, search, and heterogeneity," *Advances in Macroeconomics*, 1.

Su, Y. and J. Li (2023): "Admission Control of Double-sided Queues with Multiple Customer Types," *IEEE Transactions on Automatic Control*.

Ünver, M. U. (2010): "Dynamic kidney exchange," *The Review of Economic Studies*, 77, 372–414.

Waldinger, D. (2021): "Targeting in-kind transfers through market design: A revealed preference analysis of public housing allocation," *American Economic Review*, 111, 2660–2696.

Wang, F., J. Wang, and Z. G. Zhang (2017): "Strategic behavior and social optimization in a double-ended queue with gated policy," *Computers & Industrial Engineering*, 114, 264–273.

Wang, Z., C. Yang, and Y. Q. Zhao (2023): "Equilibrium and socially optimal strategies of a double-sided queueing system with two-mass point matching time," *Quality Technology & Quantitative Management*, 20, 89–112.

Zenios, S. A. (1999): "Modeling the transplant waiting list: A queueing model with reneging," *Queueing systems*, 31, 239–251.

A Omitted proofs

A.1 Omitted proofs from Section 3

Preliminaries We start by introducing a key piece of notation used throughout the proofs in this and the following sections. Recall that in the model of Section 3, the value of increasing the queue size from k to $k+1$, is given by

$$\mathrm{M}(k) = H_c \frac{1 + x^{k+1} - 2l_c + 2x \sum_{i=0}^{k-1} (x^k - x^i)}{(2k+1)(2k+3)} = H_c \frac{\mathrm{N}(k)}{(2k+1)(2k+3)}, \tag{A.1}$$

where the first equality repeats Equation 8 for ease of reference, and the second equality defines the function $\mathrm{N}(k)$ as the numerator of the function M evaluated at k, when the model parameters are x, l_c. As Equation A.1 makes explicit, the sign of $\mathrm{M}(k)$ equals that of $\mathrm{N}(k)$, a property that we exploit repeatedly in our arguments.

Proof of Lemma 2. The proof follows from showing the numerator $\mathrm{N}(k)$ of the marginal value in Equation A.1 is decreasing in k. Rewriting the numerator as

$$\mathrm{N}(k) = 1 + x^{k+1} - 2l_c + 2x\sum_{i=0}^{k-1}(x^k - x^i) = 1 + x^{k+1} - 2l_c + 2x\left(kx^k - \frac{1-x^k}{1-x}\right), \tag{A.2}$$

we show that for all $k \geq 0$,

$$x^{k+2} + 2x\left[(k+1)\,x^{k+1} - \frac{1-x^{k+1}}{1-x}\right] < x^{k+1} + 2x\left[kx^k - \frac{1-x^k}{1-x}\right],$$

which follows from noting $x \in (0,1)$ implies

$$(x-1)\left[x^{k+1} + 2x(k+1)\right] < 0.$$

To prove part (a), note the sign of $\mathrm{M}(k)$ is determined by the sign of its numerator, $\mathrm{N}(k)$. Because $\mathrm{N}(k)$ is decreasing, $\mathrm{M}(k+1) < 0$ whenever $\mathrm{M}(k) < 0$.

To prove part (b), note

$$\mathrm{M}(k+1) = H_c\frac{\mathrm{N}(k+1)}{(2k+5)(2k+3)} \leq H_c\frac{\mathrm{N}(k)}{(2k+5)(2k+3)} = \frac{2k+1}{2k+5}\mathrm{M}(k),$$

where the inequality follows because $\mathrm{N}(k+1) < \mathrm{N}(k)$. If $\mathrm{M}(k) \geq 0$, the result follows from noting $(2k+1)/(2k+5) < 1$. □

We now prove the claims made in Remark 4 concerning the comparative statics of K^* as a function of x (i.e., fixing l_c). The proof of Remark 4 relies on Lemma A.1, which establishes properties of the numerator $\mathrm{N}(k)$ as a function of x (cf. Equation A.2). For this reason, below, we write $\mathrm{N}(k; x)$ to emphasize the dependence of N on x:

Lemma A.1. *$\mathrm{N}(k; x)$ is strictly decreasing in x when $k = 0$ and U-shaped in x when $k > 0$.*

Proof of Lemma A.1. For $k = 0$, note

$$N(0; x) = 1 - 2l_c - x,$$

which is decreasing in x. Consider now $k > 0$. Note

$$(1-x)^2 \frac{\partial}{\partial x} \mathrm{N}(k;x) = -2 + (3+5k+2k^2)x^k - 2(1+4k+2k^2)x^{k+1} + (1+3k+2k^2)x^{k+2} \equiv \tilde{\mathrm{N}}_1.$$

Then,

$$\frac{\partial}{\partial x} \tilde{\mathrm{N}}_1 = -x^{k-1}(k+1)(1-x)\left((k+2)(2k+1)x - k(2k+3)\right).$$

The latter is positive and then negative in $x \in (0,1)$ when $k > 0$. It follows that $\tilde{\mathrm{N}}_1$ is first increasing and then increasing in x when $k > 0$. Note $\tilde{\mathrm{N}}_1(k;1) = 1$ and $\tilde{\mathrm{N}}_1(k;0) = -2$ when $k > 0$. Thus, $\tilde{\mathrm{N}}_1$ is first negative and then positive in $x \in (0,1)$. It follows that $\mathrm{N}(k;\cdot)$ is first decreasing and then increasing in $x \in (0,1)$. □

Proof of Remark 4. Define $K^\circ(x)$ to be the unique solution to

$$\mathrm{N}(k;x) = 0,$$

where, like in the proof of Lemma A.1, we write $\mathrm{N}(k;x)$ to emphasize the dependence of N on x. Note $K^* = \lceil K^\circ \rceil$. Below, we show the properties in Remark 4 hold for $K^\circ(x)$, which is enough to show they hold for K^*.

Consider first the case $l_c \geq 1/2$. Then, $\mathrm{N}(k;0) = 1 - 2l_c \leq 0$. Instead, $\mathrm{N}(k;1) = 2(1-l_c) \geq 0$. Except for $l_c = 1/2$, a unique $x^*(k)$ exists such that $\mathrm{N}(k;x^*(k)) = 0$ (cf. Lemma A.1). Furthermore, the derivative of N with respect to x must be positive at $x = x^*(k)$. By definition, $x^*(K^\circ(x)) = x$, that is, $x^*(k) = (K^\circ)^{-1}(k)$. It follows that the derivative of N with respect to x is positive when $k = K^\circ(x)$, and thus, $K^\circ(x)$ is non-decreasing by the implicit function theorem.[21]

Consider now the case in which $l_c < 1/2$. Then, $\mathrm{N}(k;0) > 0$ and $\mathrm{N}(k;1) > 0$. Lemma A.1 implies $\mathrm{N}(k;x)$ has two zero points in x, where the derivative at one point is negative and the other is positive. The result follows. □

A.2 Omitted proofs from Section 4

Proof of Proposition 1. Recall from Appendix A.1 that the sign of M coincides with that of N, so analyzing the sign of N at $(\log l_c / \log\ x) - 1$ is enough to establish the sign of M at $(\log l_c / \log\ x) - 1$. To make the dependence on l_c and x evident, define

$$\breve{\mathrm{N}}(l_c, x) \equiv \mathrm{N}\left(\frac{\log l_c}{\log x} - 1\right) = \left(2\frac{\log l_c}{\log x} - 1\right) l_c - \frac{2x - 2l_c}{1-x} - (2l_c - 1).$$

[21] Note that $\partial \mathrm{N}/\partial x|_{k=K^\circ(x)} = \partial \mathrm{N}/\partial x|_{x=x^*(k)}$.

Note $\breve{\mathrm{N}}_{11}(l_c, x)$ (the second derivative of $\breve{\mathrm{N}}$ against l_c) is negative for $l_c \in (0, 1)$. Furthermore,

$$\lim_{l_c \to 0} \breve{\mathrm{N}}(l_c, x) = \frac{1 - 3x}{1 - x} \text{ and } \breve{\mathrm{N}}(1, x) = 0.$$

It follows that when $x \leq 1/3$, $\mathrm{N}(\log l_c / \log x - 1) \geq 0$. Instead, when $x > 1/3$, a unique solution in l_c exists such that $\mathrm{N}(\log l_c / \log x - 1) = 0$—denote it $\hat{l}(x)$—such that $\mathrm{N}(\log l_c / \log x - 1) \geq 0$ if and only if $l_c \geq \hat{l}(x)$.

To see $\hat{l}(x) \leq x^2$, note $\breve{\mathrm{N}}(x^2, x) = (1 - x)^2$. Furthermore, to see that when $x \geq 1/2$, we have that $\hat{l}(x) \geq x^3$, note that $\breve{\mathrm{N}}(x^3, x) = 1 - 2x - 2x^2 + 3x^3 \leq 0$ for $x \geq 1/2$. □

Proof of Proposition 2. Part (a) is straightforward: $x < l_c$ implies $k^* = 0$ by Equation 3. Instead, $l_c < (1 + x)/2$ implies $K^* > 0$ because $\mathrm{N}(0) = 1 - 2l_c + x$.

To show part (b), we show that if $x \leq 1/3$, then $\mathrm{M}(\log l_c / \log x) > 0$, and hence, $\mathrm{M}(k^*) \geq \mathrm{M}(\log l_c / \log x) > 0$ by Equation 3. By Equation 9, this inequality implies the planner wants to form a queue of length at least $k^* + 1$, that is, $K^* > k^*$.

By Equation 8, the numerator of $\mathrm{M}(\cdot)$ evaluated at $k = \log l_c / \log x$ is

$$\begin{aligned} \mathrm{N}\left(\frac{\log l_c}{\log x}\right) &= 1 - \frac{2x}{1 - x} + l_c \left(2x \frac{\log l_c}{\log x} + \frac{2x}{1 - x} + x - 2\right) \qquad \text{(A.3)} \\ &\geq 1 - \frac{2x}{1 - x} + l_c \left(2x + \frac{2x}{1 - x} + x - 2\right) \geq 1 - \frac{2x}{1 - x} + \frac{1}{3}\left(2x + \frac{2x}{1 - x} + x - 2\right) > 0, \end{aligned}$$

where the first line follows from $x^{\frac{\log l_c}{\log x}} = l_c$, the first inequality follows from $x \geq l_c$, and the last one from the observation that the term in brackets in the third expression is negative for $x \leq 1/3$ and $x \in [l_c, 1/3]$, and the final inequality follows from algebra. Because the sign of M coincides with that of its numerator N, we arrive at the desired conclusion.

We conclude the proof by showing part (c) holds. For $n \geq 3$ and $x > 1/2$, and each $x^{n+1} < l_c \leq x^n$, note $k^* = n$ and

$$\mathrm{N}(n - 1) = 1 - 2l_c + (2n - 1)x^n - 2x\frac{1 - x^{n-1}}{1 - x} < 1 + (2n - 1)x^n - 2x^{n-1} - 2x\frac{1 - x^{n-1}}{1 - x} \equiv \widetilde{\mathrm{N}}(n - 1),$$

where the inequality follows from $l_c > x^{n+1}$. Moreover,

$$\widetilde{\mathrm{N}}(n - 1) - \widetilde{\mathrm{N}}(n) = x^n \left[2n + 1 - x(2n + 3) + 2x^2\right] = 2x^n(x - 1)(x - (n + 1/2)) \geq 0,$$

whenever $x \in (1/2, 1]$. Thus, it suffices to show $\widetilde{\mathrm{N}}(2) = 1 + 3x^2 - 2x - 2x^4 < 0$, and this holds for $x \in (1/2, 1]$. We conclude that for $n \geq 3$, $x > 1/2$, and $x^{n+1} < l_c \leq x^n$, $K^* \leq n - 1 < k^*$, completing the proof. □

Claim A.1. *Fix l_c. Then, $\hat{x}(l_c)$ exists such that for all $x \geq \hat{x}(l_c)$,* $\mathrm{M}(k^* - 1) < 0$.

Proof. It suffices to show $\lim_{x\to 1} \mathrm{N}(\log l_c/\log x - 2) < 0$. Recall that

$$\mathrm{N}\left(\frac{\log l_c}{\log x} - 2\right) = 1 - 2l_c + \frac{l_c}{x} + 2((\frac{\log l_c}{\log x} - 2)\frac{l_c}{x} - \frac{x - \frac{l_c}{x}}{1-x}).$$

Noting $\lim_{x\to 1} \frac{\log l_c}{\log x} = -\infty$ yields the result. □

Proof of Proposition 3. Recall the sign of $\mathrm{M}(k)$ is the same as the sign of its numerator $\mathrm{N}(k)$. Note that

$$\lim_{k\to\infty} \mathrm{N}(k) = \lim_{k\to\infty} 1 - 2l_c + x^{k+1} + 2x\sum_{i=0}^{k-1}(x^k - x^i) = 1 - 2l_c - 2\frac{x}{1-x}.$$

Hence, if the above condition holds, $\mathrm{M}(k) > 0$ for all k, and therefore $K^* = \infty$. □

A.3 Omitted proofs from Section 4.1

Proof of Proposition 4. Assume $n \leq \overline{K}^* \leq n+1$; that is

$$n \leq \sqrt{\frac{2(H-L)}{c}} \leq n+1.$$

Note this implies

$$n^2 \leq \frac{2(H-L)}{c} \leq (n+1)^2$$
$$\frac{n^2}{2} \leq \overline{k}^* \leq \frac{(n+1)^2}{2}.$$

When $n \geq 3$, it follows that $n+1 < n^2/2$, and hence, $\overline{K}^* < \overline{k}^*$. Moreover, suppose $\overline{K}^* = 2$, that is, $2 \leq \sqrt{2(H-L)/c} < 3$. Then, $\overline{k}^* \geq 2$. The argument in the main text implies that when $\overline{k}^* = 1$, then $\overline{K}^* = 1$ as well. □

A.4 Omitted proofs from Section 5.1

Proof of Proposition 5. The argument after the statement of Proposition 5 shows that if $k_A + k_B$ is even,

$$\hat{\mathrm{W}}(k_A, k_B) \leq \mathrm{W}\left(\frac{k_A + k_B}{2}, l_c^P\right) \leq \mathrm{W}\left(K^*\left(l_c^P\right), l_c^P\right).$$

We complete here the proof by considering the case in which $k_A + k_B$ is odd.

First, note it is without loss of generality to focus on $k_B = k_A + 1$, because $\hat{\mathrm{W}}(k_A, k_B)$ is symmetric *and* $\hat{\mathrm{W}}(k_A, k_B) \leq \hat{\mathrm{W}}(k_A + 1, k_B - 1)$ whenever $k_B > k_A + 1$ by convexity of x^k. Second, we argue $\hat{\mathrm{W}}(k,k+1) \geq \hat{\mathrm{W}}(k,k)$ implies $\hat{\mathrm{W}}(k+1,k+1) \geq \hat{\mathrm{W}}(k,k+1)$, so that we can continue to apply the argument in the main text to argue $k_A = k_B = K^*(l_c^P)$ is the optimal queue size. This second point follows from noting

$$(2k+3)\left[\hat{\mathrm{W}}(k+1,k+1) - \hat{\mathrm{W}}(k,k+1)\right] = (2k+1)\left[\hat{\mathrm{W}}(k,k+1) - \hat{\mathrm{W}}(k,k)\right],$$

completing the proof. □

Pure flow costs We include here the derivations of the planner's objective for the case in which the agents only incur in flow costs when they queue. Denote by $\overline{k}^*(L)$ and $\overline{K}^*(L)$ the equilibrium and socially optimal queue sizes derived in Section 4.1, respectively. Like in Section 5.1, these objects are enough to characterize the equilibrium and socially optimal queue sizes when outside options are different across sides and agents only experience flow costs while they wait.

When the planner sets queue sizes (k_A, k_B) for each side, welfare is given by

$$\hat{\overline{\mathrm{W}}}(k_A,k_B) = \frac{1}{1+k_A+k_B}\left(\frac{L_A+L_B}{2} + \frac{H}{2}(k_A+k_B) - \frac{c}{2\lambda}\frac{(k_A(k_A+1)+k_B(k_B+1))}{2}\right),$$

where the notation $\hat{\overline{\mathrm{W}}}$ signifies the above equation is the analogue of $\overline{\mathrm{W}}$ in Section 4.1.

Because $k(k+1)$ is convex, a similar argument to the one for Proposition 5 establishes that the optimal queue sizes are the same across sides and correspond to the optimal queue size when the outside option has value $L^P = (L_A+L_B)/2$ for both sides. That is, we have that $\overline{K}^*_A = \overline{K}^*_B = \overline{K}^*(L^P)$. By contrast, the equilibrium queue sizes are given by

$$\overline{k}^*_A = \overline{k}^*(L_A) \text{ and } \overline{k}^*_B = \overline{k}^*(L_B).$$

Proof of Proposition 6. Suppose $\overline{k}^*_A$ and $\overline{k}^*_B$ satisfy that

$$n_A \leq \overline{k}^*_A \leq n_A+1 \text{ and } n_B \leq \overline{k}^*_B \leq n_B+1.$$

Then,

$$\sqrt{n_A+n_B} \leq \overline{K}^*(L^P) \leq \sqrt{n_A+n_B+2}. \tag{A.4}$$

Letting $N = (n_A+n_B)/2$, showing $\sqrt{2N+2} \leq N+1$ is enough to conclude $\overline{K}^*(L^P) \leq N = (n_A+n_B)/2$. This holds whenever $N = (n_A+n_B)/2 > 1$.

Suppose now that $n_A + n_B \leq 2$. We consider three cases:

Case 1 $n_A + n_B = 2$ Then Equation A.4 implies $1 \le \sqrt{2} \le \overline{K}^*(L^P) \le 2$. However, in this case having $\overline{K}^*(L^P) = 2$ is not possible. To see this, note $\overline{K}^*(L^P) = 2$ implies

$$\lambda \frac{H - L_A}{c} + \lambda \frac{H - L_B}{c} = 4.$$

Furthermore, $n_A + n_B = 2$ implies either $n_A = 2$ and $n_B = 0$ or $n_A = n_B = 1$. In the first case, we obtain that

$$\lambda \frac{H - L_A}{c} + \lambda \frac{H - L_B}{c} < 3 + 1 = 4,$$

where the first part of the inequality follows from $n_A < 3$ and the second from $n_B < 1$. Instead, if $n_A = n_B = 1$,

$$\lambda \frac{H - L_A}{c} + \lambda \frac{H - L_B}{c} < 4,$$

again a contradiction.

Case 2 $n_A + n_B = 1$ In this case, only one of the sides queues, and it must be side A. Equation A.4 implies $1 \le \overline{K}^*(L^P)$, so the socially optimal queue is longer than the average equilibrium queue length.[22]

Case 3 $n_A + n_B = 0$ In this case, neither side queues and Equation A.4 implies that $\overline{K}^*(L^P) \le \sqrt{2} < 2$, so the socially optimal queue size is at most 1. This is analogous to the case in Proposition 4 in Section 4.1 when the planner is willing to have a queue of size 1 even if the agents are not willing to join.

□

A.5 Omitted proofs from Section 5.2

In Appendix A.5, we provide proofs of several statements in Section 5.2. Lemma A.2 and Corollary A.1 show fixing the opposite side's queue size, the marginal values for each side, $\hat{\mathrm{M}}_A$ and $\hat{\mathrm{M}}_B$, retain the same properties as the marginal value M in the main analysis. We use these properties to prove Proposition 7, which relies on Lemma A.3.

Define:

$$\hat{\mathrm{M}}_A(k_A, k_B) = \hat{\mathrm{W}}(k_A + 1, k_B) - \hat{\mathrm{W}}(k_A, k_B), \;\; \hat{\mathrm{M}}_B(k_A, k_B) = \hat{\mathrm{W}}(k_A, k_B + 1) - \hat{\mathrm{W}}(k_A, k_B).$$

Moreover, let $\pi^+_{k_B}$ denote $(1 - \rho)/(1 - \rho^{k_A + k_B + 2})$, denote the probability of state k_B when the planner chooses sizes $(k_A + 1, k_B)$ and the probability of state $k_B + 1$ when the planner chooses

[22] A numerical instance of this is $H = 1, L_A = 0.25, L_B = 7/16, c = 1, \lambda = 25/16$.

sizes (k_A, k_B+1). We then have the following:

$$\hat{\mathrm{M}}_A(k_A,k_B) = \frac{\pi_{k_B}\pi^+_{k_B}\rho^{k_A+k_B+1}}{1+\rho}\left[\frac{x_A^{k_A+1}}{\pi_{k_B}} + 1 - 2l_c - \rho^{k_B}\left(\rho x_A\frac{1-(\rho x_A)^{k_A}}{1-\rho x_A} + x_B\frac{1-(x_B/\rho)^{k_B}}{1-(x_B/\rho)}\right)\right],$$

$$\hat{\mathrm{M}}_B(k_A,k_B) = \frac{\pi_{k_B}\pi^+_{k_B}\rho^{k_B+1}}{1+\rho}\left[\frac{\rho}{\pi_{k_B}}\left(\frac{x_B}{\rho}\right)^{k_B+1} + \rho^{k_A}(1-2l_c) - \left(x_A\frac{1-(\rho x_A)^{k_A}}{1-\rho x_A} + \frac{x_B}{\rho}\frac{1-(x_B/\rho)^{k_B}}{1-(x_B/\rho)}\right)\right].$$

Similarly to the analysis in Section 3, define $\hat{\mathrm{N}}_A$ and $\hat{\mathrm{N}}_B$ to be the terms in square brackets in $\hat{\mathrm{M}}_A$ and $\hat{\mathrm{M}}_B$, respectively. Note these terms determine the sign of $\hat{\mathrm{M}}_A$ and $\hat{\mathrm{M}}_B$. We have the following:

Lemma A.2. *For each $k_B \geq 0$, $\hat{\mathrm{N}}_A(\cdot, k_B)$ is decreasing in k_A. Similarly, for each $k_A \geq 0$, $\hat{\mathrm{N}}_B(k_A, \cdot)$ is decreasing in k_B.*

Proof. Consider first $\hat{\mathrm{N}}_A(\cdot, k_B)$. We need to show

$$\frac{x_A^{k_A+2}}{\pi^+_{k_B}} - \rho^{k_B}\left(\rho x_A\frac{1-(\rho x_A)^{k_A+1}}{1-\rho x_A}\right) < \frac{x_A^{k_A+1}}{\pi_{k_B}} - \rho^{k_B}\left(\rho x_A\frac{1-(\rho x_A)^{k_A}}{1-\rho x_A}\right).$$

Now,

$$\frac{x_A^{k_A+1}}{\pi_{k_B}} - \frac{x_A^{k_A+2}}{\pi^+_{k_B}} = x_A^{k_A+1}\left(\sum_{s=0}^{k_A+k_B}\rho^s - x_A\sum_{s=0}^{k_A+k_B+1}\rho^s\right) = x_A^{k_A+1}\left((1-x_A)\sum_{s=0}^{k_A+k_B}\rho^s - x_A\rho^{k_A+k_B+1}\right).$$

Moreover,

$$\rho^{k_B}\rho x_A\left(\sum_{s=0}^{k_A}(\rho x_A)^s - \sum_{s=0}^{k_A-1}(\rho x_A)^s\right) = \rho^{k_B}(\rho x_A)^{k_A+1}$$

Adding up both terms we get:

$$\begin{aligned} &x_A^{k_A+1}\left((1-x_A)\sum_{s=0}^{k_A+k_B}\rho^s - x_A\rho^{k_A+k_B+1}\right) + \rho^{k_B}(\rho x_A)^{k_A+1} \\ &= x_A^{k_A+1}(1-x_A)\frac{1-\rho^{k_A+k_B+1}}{1-\rho} + \rho^{k_A+k_B+1}x_A^{k_A+1}(1-x_A) > 0, \end{aligned}$$

so that $\hat{\mathrm{N}}_A(k_A+1,k_B) < \hat{\mathrm{N}}_A(k_A,k_B)$. Consider now $\hat{\mathrm{N}}_B(k_A,\cdot)$. We need to show

$$\frac{\rho}{\pi^+_{k_B}}\left(\frac{x_B}{\rho}\right)^{k_B+2} - \frac{x_B}{\rho}\frac{1-(x_B/\rho)^{k_B+1}}{1-x_B/\rho} < \frac{\rho}{\pi_{k_B}}\left(\frac{x_B}{\rho}\right)^{k_B+1} - \frac{x_B}{\rho}\frac{1-(x_B/\rho)^{k_B}}{1-x_B/\rho}.$$

Now,

$$\frac{\rho}{\pi_{k_B}}\left(\frac{x_B}{\rho}\right)^{k_B+1}-\frac{\rho}{\pi^{+}_{k_B}}\left(\frac{x_B}{\rho}\right)^{k_B+2}=\rho\left(\frac{x_B}{\rho}\right)^{k_B+1}\left(\left(1-\frac{x_B}{\rho}\right)\sum_{s=0}^{k_A+k_B}\rho^s-\frac{x_B}{\rho}\rho^{k_A+k_B+1}\right).$$

Moreover,

$$\left(\frac{x_B}{\rho}\right)\left(\sum_{s=0}^{k_B}(x_B/\rho)^s-\sum_{s=0}^{k_B-1}(x_B/\rho)^s\right)=(x_B/\rho)^{k_B+1}.$$

Adding up both terms, we get

$$\left(\frac{x_B}{\rho}\right)^{k_B+1}\left[\rho\left(1-\frac{x_B}{\rho}\right)\frac{1-\rho^{k_A+k_B+1}}{1-\rho}+1-x_B\rho^{k_A+k_B+1}\right]>0,$$

so that $\hat{\mathrm{N}}_B(k_A,k_B+1)<\hat{\mathrm{N}}_B(k_A,k_B)$, completing the proof. □

Corollary A.1. *The following holds for each $k_B\in\mathbb{N}_0$:*

1. *If $\hat{\mathrm{M}}_A(k_A,k_B)<0$, then for all $k'_A>k_A$, $\hat{\mathrm{M}}_A(k'_A,k_B)<0$,*
2. *If $\hat{\mathrm{M}}_A(k_A,k_B)\geq 0$, then $\hat{\mathrm{M}}_A(k_A+1,k_B)\leq\hat{\mathrm{M}}_A(k_A,k_B)$.*

By appropriately changing the roles of sides A and B, Corollary A.1 also holds for $\hat{\mathrm{M}}_B(k_A,\cdot)$.

Proof of Corollary A.1. Part 1 follows immediately from Lemma A.2 and the observation that the sign of $\hat{\mathrm{M}}_A(\cdot,k_B)$ coincides with that of $\hat{\mathrm{N}}_A(\cdot,k_B)$.

To see part 2 holds, define $\pi^{++}_{k_B}=(1-\rho)/(1-\rho^{k_A+k_B+3})$. Then,

$$\hat{\mathrm{M}}_A(k_A+1,k_B)=\frac{\pi^{++}_{k_B}\pi^{+}_{k_B}\rho^{k_A+k_B+2}}{1+\rho}\hat{\mathrm{N}}_A(k_A+1,k_B)$$
$$\leq\frac{\pi^{++}_{k_B}\pi^{+}_{k_B}\rho^{k_A+k_B+2}}{1+\rho}\hat{\mathrm{N}}_A(k_A,k_B)=\rho\frac{\pi^{++}_{k_B}}{\pi_{k_B}}\hat{\mathrm{M}}_A(k_A,k_B)=\rho\frac{1-\rho^{k_A+k_B+1}}{1-\rho^{k_A+k_B+3}}\hat{\mathrm{M}}_A(k_A,k_B)\leq\hat{\mathrm{M}}_A(k_A,k_B),$$

where the first inequality follows from Lemma A.2 and the second inequality from the assumption that $\hat{\mathrm{M}}_A(k_A,k_B)\geq 0$. □

Proof of Proposition 7 Lemma A.3 is a key intermediate step in the proof of Proposition 7:

Lemma A.3. *If*

$$1-2l_c-\frac{\rho x_A}{1-\rho x_A}-x_B\frac{\frac{\ln x_B}{\ln\rho}-1}{1-x_B/\rho}>0,$$

$K^*_A(k_B)=+\infty$ is optimal.

In words, when the condition in Lemma A.3 holds, the planner finds it optimal to holds an infinite queue on side A regardless of the queue size on side B.

Proof of Lemma A.3. Note that for each k_B,

$$\lim_{k_A \to +\infty} \hat{\text{N}}_A(k_A, k_B) = 1 - 2l_c - \rho^{k_B} \frac{\rho x_A}{1 - \rho x_A} - x_B \frac{\rho^{k_B} - x_B^{k_B}}{1 - x_B/\rho}.$$

Define the function $G(k_B) = \rho^{k_B} - x_B^{k_B}$. Note $G(0) = \lim_{k_B \to \infty} G(k_B) = 0$ and $G(1) > 0$. Furthermore, its derivative is given by

$$G'(k_B) = \rho^{k_B} \ln \rho - x_B^{k_B} \ln x_B.$$

Note

$$k_B^\circ = \frac{\ln\left(\frac{\ln x_B}{\ln \rho}\right)}{\ln(\rho/x_B)},$$

is such that $G'(k_B^\circ) = 0$. Furthermore, note $\lim_{k_B \to +\infty} G'(k_B) = 0$. However, if $G'(k_B) < 0$, then $G'(k_B') < 0$ for $k_B' > k_B$. To see this, note the sign of $G'(k_B)$ coincides with the sign of

$$\left(\frac{\rho}{x_B}\right)^{k_B} \ln \rho - \ln x_B, \tag{A.5}$$

which is decreasing in k_B as $\rho/x_B > 1$ and $\ln \rho < 0$. We use this in what follows to argue that k_B° is the maximizer of G.

Note

$$G''(k_B) = \rho^{k_B} (\ln \rho)^2 - x_B^{k_B} (\ln x_B)^2,$$

which is negative at $k_B = 0$, goes to 0 as $k_B \to \infty$, and is zero at $k_B^{\circ\circ} > k_B^\circ$, where

$$k_B^{\circ\circ} = \frac{\ln\left(\frac{(\ln x_B)^2}{(\ln \rho)^2}\right)}{\ln(\rho/x_B)}.$$

It follows that G' is decreasing on $[0, k_B^{\circ\circ}]$ and then increasing on $[k_B^{\circ\circ}, +\infty)$. Moreover, it remains below (and approaches from below) 0 as $k_B \to +\infty$. We conclude k_B° is the maximizer of G.

Note then that

$$\lim_{k_A\to\infty} \hat{\mathrm{N}}_A(k_A,k_B) \geq 1-2l_c-\frac{\rho x_A}{1-\rho x_A}-x_B\frac{\rho^{k_B^\circ}-x_B^{k_B^\circ}}{1-x_B/\rho} = 1-2l_c-\frac{\rho x_A}{1-\rho x_A}-x_B^{k_B^\circ+1}\frac{\frac{\ln x_B}{\ln\rho}-1}{1-x_B/\rho}$$

$$\geq 1-2l_c-\frac{\rho x_A}{1-\rho x_A}-x_B\frac{\frac{\ln x_B}{\ln\rho}-1}{1-x_B/\rho}. \tag{A.6}$$

Thus, if

$$1-2l_c-\frac{\rho x_A}{1-\rho x_A}-x_B\frac{\frac{\ln x_B}{\ln\rho}-1}{1-x_B/\rho} > 0, \tag{A.7}$$

$K_A^*(k_B)=+\infty$ for each k_B concluding the proof. □

Proof of Proposition 7. Define $\hat{\mathrm{N}}_\infty(k_B) = \lim_{k_A\to\infty}\hat{\mathrm{N}}_B(k_A,k_B)$. Manipulating the expression for $\hat{\mathrm{N}}_\infty(k_B)$ yields that $\hat{K}_B^*\geq 0$ defined by

$$\left(\frac{x_B}{\rho}\right)^{\hat{K}_B^*+1} \leq \frac{2(\mu-\lambda)(\lambda+r)}{(\mu+r-\lambda)\mu} \leq \left(\frac{x_B}{\rho}\right)^{\hat{K}_B^*} \tag{A.8}$$

is the optimal side B queue size given that the side A queue is unbounded. Interestingly, the side B optimal queue size is independent of the value of the outside option, because the probability of the event in which an agent on side B takes their outside option—agents on side A never do—is independent of the queue size on side B when $k_A\to\infty$ and equal to $1-\rho$.

Now, when $\mu\geq 2\lambda$,

$$\frac{2(\mu-\lambda)(\lambda+r)}{(\mu+r-\lambda)\mu}\geq 1,$$

and hence $\hat{K}_B^*=0$. When $\mu<\min\{2\lambda,\lambda+r\}$,[23] one can establish numerically that it is not possible that Equation A.7 holds and $\hat{K}_B^*\geq 1$ solves Equation A.8. That is, for any tuple of parameters (λ,μ,r,l_c) such that Equation A.7 holds, Equation A.8 fails to hold at $\hat{K}_B^*\geq 1$, and vice versa. □

[23] Recall we are assuming that $x_B<\rho$ and hence $\mu<\lambda+r$.