Optimal Allocation via Waitlists: Simplicity through Information Design*

Itai Ashlagi, Faidra Monachou, Afshin Nikzad†

June 12, 2023

Abstract

We study nonmonetary markets where objects that arrive over time are allocated to unit-demand agents with private types, such as in the allocation of public housing or deceased-donor organs. An agent's value for an object is supermodular in her type and the object quality, and her payoff is her value minus her waiting cost. The social planner's objective is a weighted sum of allocative efficiency (i.e., the sum of values) and welfare (i.e., the sum of payoffs). We identify optimal mechanisms in the class of direct-revelation mechanisms. When the social planner can design the information disclosed to the agents about the objects, the optimal mechanism has a simple implementation: a first-come first-served waitlist with deferrals. In this implementation, the object qualities are partitioned into intervals; only the interval containing the object quality is disclosed to agents. When the planner places a higher weight on welfare, optimal disclosure policies become coarser.

*A preliminary version of this paper appeared in the Proceedings of the 2021 ACM Conference on Economics and Computation (EC'21). The authors thank Gabriel Carroll, Yeon-Koo Che, Haluk Ergin, Marina Halac, Yash Kanoria, Andreas Kleiner, Jonathan Libgober, Vahideh Manshadi, Phil Reny, Joe Root, Andy Skrzypacz, Tayfun Sönmez, Philipp Strack, Guofu Tan, Olivier Tercieux, Rakesh Vohra, and numerous seminar audiences, commentators, and anonymous referees.

†Ashlagi: MS&E, Stanford University, iashlagi@stanford.edu; Monachou: CMSA, Harvard University, faidra.monachou@yale.edu; Nikzad: Economics, University of Southern California, afshin.nikzad@usc.edu.

1 Introduction

Many scarce resources, such as public housing or deceased-donor organs, are allocated through waitlists to agents with privately known preferences. For example, in many countries deceased-donor organs are allocated through a priority-based waitlist mechanism. Despite the severe shortage of organs, patients on the waitlist may choose to wait for higher-quality organs: those who value quality more, such as some of the younger patients, are willing to wait longer (Agarwal et al., 2021). Waiting is also costly, since the patients on the waitlist need to undergo dialysis and also may become too sick to receive a transplant. This leads to a tradeoff between lower waiting times and higher allocative efficiency.

This paper is broadly concerned with how to optimally resolve this tradeoff in nonmonetary markets where agents and objects arrive over time and the agents' preferences over the objects are unobservable. Are there simple waitlist implementations of the optimal solution without explicitly eliciting preferences? In what way do optimal solutions change based on how much the planner weighs allocative efficiency over waiting times?

We address these questions in a setting where the objects differ in terms of *quality* and the agents vertically differ in their preferences for quality. For example in the context of deceased-donor kidneys, the Kidney Donor Profile Index (KDPI) is a standard measure for organ quality, which summarizes clinical parameters and demographics into a single number (OPTN, 2022).[1]

The planner's objective in our setting is a weighted sum of *allocative e ciency* and *social welfare*, where the former is the surplus generated from the allocation (ignoring waiting times), and the latter is the average of the agents' payoffs, accounting for waiting times.[2] As an agent's preference for quality is her private information, we take a mechanism design approach.

First, we identify optimal mechanisms in the space of direct-revelation mechanisms that elicit the agents' preferences and match them to objects over time. We then show that the mechanism that we identify is implementable by a simple first-come first-served (FCFS) waitlist with deferrals, if the planner can design what information is disclosed to the agents about object qualities. The information disclosed about each object is an interval containing the object quality. These intervals essentially partition the quality space. We show that these intervals capture the tradeoff between allocative efficiency and social welfare in an intuitive way: on the Pareto frontier defined by direct-revelation mechanisms, the mechanisms that achieve higher social welfare use *coarser* intervals to pool object types. This means implementing outcomes with higher welfare requires less informative disclosure policies.

More precisely, our main setup features agents and objects that arrive over time at possibly

[1] We extend our findings to scenarios where objects are horizontally differentiated in Online Appendix V.

[2] For example, in the allocation of deceased-donor organs, while waiting times are identified as an important factor that should be taken into account (OPTN, 2015), allocative efficiency is also one of the criteria of interest. The Organ Procurement and Transplantation Network added a reserve policy in 2014 based on which 20% of the highest-quality kidneys are reserved for adults with the highest 20% post-transplant survival score (OPTN, 2014).

different rates. Agents have private types that capture their preference for quality. An agent's value for an object is a supermodular function of the agent's type and the object's quality.[3] An agent's payoff is her value for the (expected) object quality that she is allocated minus her waiting cost. The planner's objective is to maximize a weighted sum of allocative efficiency and social welfare. Due to the supermodularity of the utility function, a positive assortative assignment maximizes allocative efficiency, but not necessarily social welfare. We first ask what mechanisms maximize the planner's objective in the space of all *steady-state* direct revelation mechanisms.

We focus on *large* markets to answer this question. This allows us to optimize over the space of all direct revelation mechanisms. Related work in the dynamic matching literature restricts attention to optimizing over special classes of mechanisms (Su and Zenios, 2006; Bloch and Cantala, 2017; Leshno, 2022) or assumes that agents' preferences are observable (Baccara et al., 2020; Akbarpour et al., 2020; Che and Tercieux, 2020).

The optimal mechanism is characterized by a finite set of disjoint queues. The space of object qualities is partitioned into intervals, and the objects belonging to same interval are sent to the same queue. Every agent, upon arrival, is assigned to a queue based on her reported type. Within each queue, objects are allocated to agents according to a first-in first-out (FIFO) scheme: when an object becomes available, it is assigned to the agent with the earliest arrival time present in the queue; the assignment is definitive and cannot be declined.[4]

In numerous applications there is a single waitlist and agents can *decline* offers and maintain their position on the list. We show that offering this flexibility to agents is without loss if the planner can design the information disclosed about objects. This allows for an implementation of the optimal mechanism using a single FCFS waitlist with deferrals.

In a FCFS waitlist with deferrals, every arriving object is offered to the agents on the waitlist in the order of their arrival time, and the object is allocated to the first agent who accepts the offer. Hence, agents who value quality more would wait longer. Intuitively, waiting costs act as prices to clear the market, with higher-quality objects having higher prices. When object qualities are fully disclosed, higher-quality objects are allocated to agents with a higher value for quality. It follows that full disclosure of quality maximizes allocative efficiency. However, it may not maximize social welfare, because individuals do not internalize the negative externalities of waiting on others. An information disclosure policy may allow for changing the waiting times such that efficiency gains due to lower waiting times offset losses from the non-assortative assignment of objects to agents.

We show that a FCFS waitlist with deferrals can attain the same objective value as the optimal mechanism when it uses a disclosure policy that partitions the object qualities into intervals, and discloses only the interval containing the object quality to the agents. This is the same set of

[3]The supermodularity of the utility function facilitates capturing scenarios that cannot be easily captured otherwise, such as scenarios where an agent, after observing the allocation mechanism, can change her type by taking costly actions. See Example 2.2 for details.

[4]The designer has full commitment power and implements a promised (possibly probabilistic) allocation.

intervals used in the optimal direct-revelation mechanism. The challenge in proving this result is analyzing a dynamic game and accounting for the strategic behavior of agents, who reason about the decisions of the others ahead of them on the list. We show that this game has a unique equilibrium, and the planner's objective at this equilibrium equals that in the optimal disjoint queue mechanism.

In Section 5 we study how optimal mechanisms change with the planner's objective, focusing on the case of multiplicatively separable utility functions.[5] Employing a result from Nikzad (2023), we find monotone comparative statics: as the planner places more weight on social welfare, optimal mechanisms pool *larger* subsets of agent types, in the set-inclusion order. In the context of FCFS with deferrals, this means that optimal disclosure policies use coarser intervals to pool object types. Hence, there is a tradeoff between social welfare and informativeness on the Pareto frontier. Moreover, we show that this tradeoff is resolved in a simple way when there is a *severe* shortage of supply: Then, a full-disclosure (no-disclosure) policy is optimal if the planner places more (less) weight on allocative efficiency than on social welfare.[6]

For simplicity, we first consider markets where a continuum of agents and objects arrive according to constant rates, and focus on equilibria where the market composition remains at a steady state (constant) across time. In Section 6 we consider discrete markets where agents and objects arrive according to Poisson processes, and the market composition changes across time. Then, the space of direct-revelation mechanisms is much larger, since mechanisms can determine how to allocate objects based on the market composition. Yet, we show that a natural adaptation of monotone disjoint queue mechanisms attains the second-best objective value when the market grows large. In this adaptation objects are pooled in the same way as in the continuum model, but agents are assigned to queues based on the real-time queue lengths, so that each agent's expected payoff is maximized. (Thus, agents of the same type may join different queues, unlike in the continuum model.)

We build on techniques from static mechanism design settings to develop optimal mechanisms in dynamic matching markets where objects are allocated to unit-demand agents. In terms of methodology, the proof for the optimality of monotone disjoint queue mechanisms does not use classical ironing techniques (Myerson, 1981; Toikka, 2011). These techniques are not directly applicable due to the concurrence of two complicating factors: the supermodularity of the agents' values for objects and the objects' capacity constraints. Instead, we build on the work of Kleiner et al. (2021), who characterize the extreme points of a set of functions that are majorized by a reference function. On the other hand, our monotone comparative statics analysis focuses on multiplicatively separable utility functions and applies ironing. This analysis is based on a novel result that we derive for the ironing procedure, which could be of independent interest.

[5] The multiplicative-separability assumption is ubiquitous in the mechanism design literature. For example, the revenue-maximizing auctions of Myerson (1981) are characterized under this condition.

[6] It follows that restricting attention to one of these two simple policies results in a diminishingly small objective loss when the agents' arrival rate grows large.

Our main setup assumes that objects have common qualities. The insights developed here apply to settings where agents can belong to several groups and across different groups agents may assign different qualities to objects (for example, due to incompatibilities). The planner can then set quotas for assigning objects to each group and allocate objects to each group according to the mechanisms we described. These extensions are presented in the online appendix.

The tradeoff between lower waiting times and higher allocative efficiency is present in many dynamic matching markets. We analyze this tradeoff in markets where the agents' preferences for objects are unobservable. Simple queue-based mechanisms that pool adjacent object types can optimally resolve the tradeoff. An alternative design that allows agents to decline offers is a FCFS waitlist with partial disclosure of information.[7] A common way of informing agents in practice is using quality measures, such as the Kidney Donor Profile Index (KDPI) or donor age in the context of organ allocation. Coarsening or refining these measures could provide a simple way to optimally resolve the tradeoff. Coarser measures (more pooling) should be considered when there is more emphasis on low waiting times. When there is a severe shortage of supply, the planner may restrict attention to two simple policies: complete pooling and no pooling.

1.1 Related literature

This paper relates to several strands of literature. The first strand is the literature on dynamic matching and allocation of resources through queuing mechanisms. Most of the work in this area either focuses on optimizing over special classes of mechanisms (Su and Zenios, 2006; Bloch and Cantala, 2017; Doval and Szentes, 2019; Arnosti and Shi, 2020; Leshno, 2022) or assumes that the agents' preferences are observable and optimizes over general classes of mechanisms (Baccara et al., 2020; Akbarpour et al., 2020; Che and Tercieux, 2020).

Among the papers that focus on special classes of mechanisms, Su and Zenios (2006) study a class of mechanisms where there is a FCFS queue for each organ type, and the agents are assigned to these queues in an incentive-compatible way so that the sum of their discounted allocation utilities is maximized. They further impose the following condition to ensure optimality: Consider a large static market. If all agents and objects changed their type to the highest type and then objects were distributed among agents, the sum of the utilities would be smaller than if every agent changed her type to the lowest type and received an object of the lowest type.

Bloch and Cantala (2017) analyze a waitlist with no exogenous departures or discounting and with linear waiting costs. They show that agents prefer FCFS to a lottery in a private value setting, and prefer FCFS to any other mechanism in a common-value setting. Leshno (2022) analyzes a class of buffer-queue mechanisms, with the objective of maximizing allocative efficiency in an overloaded

[7] Partial information disclosure is observed, e.g., in some of the waitlists that allocate graduate housing (Stanford R&DE, 2022). In the allocation of deceased-donor kidneys as well some waitlists do not fully disclose the donor information; for example, in some European centers the results for some biopsies are disclosed only after the decision to accept or reject the organ is made (Reese et al., 2021).

system. Among other findings, he shows that service-in-random-order (SIRO) is robustly optimal. Arnosti and Shi (2020) compare various lottery formats in a setup with heterogeneous match qualities and show that using a common lottery for all objects can improve match qualities.[8]

There is also work that optimizes over general classes of mechanisms under observability assumptions. Greedy-like policies are shown to be optimal in some markets where agents are matched to each other (Baccara et al., 2020; Ashlagi et al., 2022b). Che and Tercieux (2020) consider a setting with homogeneous objects where the designer adjusts the agents' incentives for remaining in the queue through information disclosure about the queue length. They find that FCFS with no information disclosure is optimal in a rich domain. Several other papers consider information design in a single queue that serves customers on a FCFS basis (Simhon et al., 2016; Lingenbrink and Iyer, 2017; Anunrojwong et al., 2020) and identify optimal signaling schemes for recommending whether agents should join the queue or take their outside option under certain preference structures.

Other related work concerns costly signaling settings. Hartline and Roughgarden (2008) study optimal allocation of identical objects to maximize the residual surplus. Condorelli (2012) extends this to a setting with heterogeneous qualities. Both papers are concerned with multiplicatively separable utilities. Our assumption of supermodular utilities requires novel techniques compared to these prior studies. Our comparative statics result is applicable to both settings, providing insights about how the structure of optimal mechanisms varies on the Pareto frontier.

Similar to Condorelli (2012), Kleiner et al. (2021) study a static matching model with separable utilities as an application of their extreme-point characterization results; applying those results, they derive conditions under which the optimal mechanism is positive assortative or a lottery (i.e., pools all objects together). In a concurrent work, Dworczak et al. (2021) use concavification techniques to study redistribution in a static market, where the objective is maximizing a linear combination of revenue and total surplus. They show that, under certain conditions, it is optimal to augment market mechanisms with lotteries that allocate some of the objects at a reduced price. (Ortoleva et al., 2021) study the allocation of vertically differentiated objects to two groups of agents, where the utility function of one group is "more convex" than the other. Among other results, they derive insights about the structural differences between the first- and second-best solutions. In particular, second-best solutions can involve disposal of objects.

2 Setup

Consider a market in which agents and objects arrive over time at a pool, each with a flow of a constant rate. (We consider discrete markets where the arrival processes are Poisson in Section 6.) The arrival rate of agents is normalized to 1. Every object has a type that belongs to a finite set $\Omega = \{\omega_0, \omega_1, \ldots, \omega_n\}$, where $\omega_0, \ldots, \omega_n$ are distinct nonnegative reals ordered increasingly. An

[8] See also Thakral (2019).

object of type ω_i has a *higher* type than an object with type ω_j if $\omega_i > \omega_j$. We also refer to an object type as its *quality*.

Objects of type $\omega \in \Omega$ arrive at the pool at a constant rate of $F(\omega)$. The set of object types and their arrival rates are common knowledge. There is an *abundant* supply of objects of type ω_0: the arrival rate of these objects, $F(\omega_0)$, equals the arrival rate of agents, 1. We call ω_0 the *outside option* and every other member of Ω a *positive* object type. Let $\Omega_+ = \{\omega_1, \ldots, \omega_n\}$ denote the set of positive object types, and let $N = \sum_{\omega \in \Omega_+} F(\omega)$ be the sum of their arrival rates. We assume that there is no excess supply, i.e., $N \leq 1$.[9]

Every agent has a private type θ that is independently and identically distributed (i.i.d.) across all agents according to a continuous CDF G with PDF g and support $\Theta = [0, \bar{\theta}]$.[10] The distribution G is common knowledge. The utility that an agent derives from a random object depends on the agent's type and the object's expected quality. Formally, the utility that an agent of type θ assigns to a lottery over objects is $u(\theta, \bar{\omega})$, where $\bar{\omega}$ denotes the expected object type of the lottery and $u : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$. We thus interpret θ as the agent's preference for quality.[11]

Mechanisms. A *mechanism*, typically denoted by M, elicits from every agent her type upon her arrival and assigns her an object at a time (possibly) after her arrival, as detailed below.

Every agent reports her type to the mechanism upon arrival to the pool. The mechanism then assigns an *allocation timeline* $P_M(\theta)$ to the agent based on her reported type θ. An allocation timeline $P_M(\theta)$, also denoted by P_M^θ, is a joint probability distribution over object type and waiting time: for $\omega \in \Omega$, $P_M^\theta(t, \omega)$ is the probability that the agent is assigned an object with type ω at most t units of time after her report time. We focus on the set of mechanisms in which the pool remains at a *steady state*; that is, the density of agents of type θ and waiting time s present in the pool at any time remains the same across time, for every $\theta \in \Theta$ and $s \geq 0$. Therefore, the allocation timeline assigned to an agent is assumed to be a function only of her reported type, and not of the composition of the pool. In Section 6 we consider a discrete setup where the steady state assumption is relaxed.[12] The function $P_M(\cdot)$ is called the mechanism's *allocation rule*.

With slight abuse of notation, we use $P_M^\omega(\theta)$ to denote the probability that an agent of type θ is allocated an object of type ω in the mechanism M. We say M is *feasible* if $\int_{\theta \in \Theta} P_M^\omega(\theta) \mathrm{d}\theta \leq F(\omega)$ for all $w \in \Omega_+$. We restrict attention to feasible mechanisms throughout.

An agent who waits t units of time incurs a waiting cost of $c(t)$ where $c : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly

[9] The findings of this section and Section 3 hold identically when there is excess supply (i.e., $N > 1$). We use the assumption $N \leq 1$ in Section 4, to establish the uniqueness of equilibrium.

[10] In the case of a continuum of agents here, this is equivalent to assuming that the types of agents arriving at every time are distributed according to G.

[11] We extend our setup to allow for horizontal differentiation in Online Appendix V.

[12] The steady-state assumption made in the discrete setup is the usual one in stochastic processes: the existence of a steady-state distribution. There, we see that fluctuations from the average case of the stochastic process occur with a probability that approaches zero as the market grows large. Hence, when the market is infinitely large such fluctuations occur with probability zero, which is the steady-state assumption that we make in the continuum setting.

increasing continuous function that approaches infinity as t does, and $c(0) = 0$.[13] Let $c_M(\theta)$ denote the expected *waiting cost* of an agent who reports a type θ to a mechanism M; i.e., $c_M(\theta) = \mathsf{E}_t[c(t)]$ where the random variable t is the waiting time of the agent under the allocation timeline P_M^θ. Similarly, let $u_M(\theta^0;\theta)$ be the *expected utility* of an agent of type θ from reporting type θ^0 to M; i.e., $u_M(\theta^0;\theta) = \mathsf{E}_\omega[u(\theta,\omega)]$, where the random variable ω is the object type assigned to the agent under the allocation timeline $P_M^{\theta^0}$.

The *payo* of an agent of type θ from reporting type θ^0 to M is $v_M(\theta^0;\theta) = u_M(\theta^0;\theta)$ $c_M(\theta^0)$. If the agent is not matched after t units of time following her arrival, then her *continuation payo* is

$$\frac{1}{\mathsf{P}_{(\omega,s)\ P_M^{\theta^0}}[s>t]}\mathsf{E}_{(\omega,s)\ P_M^{\theta^0}}[u(\theta,\omega) \quad (c(s) \quad c(t))|s>t]$$

if the allocation timeline $P_M^{\theta^0}$ allocates an object with a positive probability after time t; otherwise, the continuation payoff is 0 by definition. Here the random variables w, s respectively denote the object type allocated to the agent and the time of allocation under $P_M^{\theta^0}$.

A mechanism is *incentive-compatible* (IC) if $v_M(\theta;\theta)$ $v_M(\theta^0;\theta)$ for every $\theta, \theta^0 \in \Theta$; i.e., the payoff of every participating agent is maximized when she reports her true type to the mechanism. A mechanism is *individually rational* (IR) if the expected continuation payoff of every unassigned agent is positive at any time in that mechanism.

The Planner's Objective. The planner's objective is maximizing a weighted combination of *allocative e ciency* and *social welfare*, as formally defined next. The *allocative e ciency* of a mechanism M is defined by $E_M = \mathsf{E}_{\theta\ G}[u_M(\theta;\theta)]$. The *social welfare* of M is defined by $W_M = \mathsf{E}_{\theta\ G}[v_M(\theta;\theta)]$.[14] We often refer to allocative efficiency as *e ciency* and social welfare as *welfare*, for brevity. The planner's objective is choosing an individually rational and incentive-compatible mechanism M that maximizes a weighted sum of welfare and efficiency

$$\lambda_W W_M + \lambda_E E_M,$$

where the weights $\lambda_W, \lambda_E \in \mathsf{R}_+$ are respectively the planner's *Pareto weights* on welfare and efficiency. Such a mechanism is called *optimal* and is typically denoted by M .

The Pareto Frontier. Another notion relevant to optimal mechanisms that we will study is the Pareto frontier. Formally, for an individually rational and incentive-compatible mechanism M, we say that M *generates* the point (E_M, W_M), where recall that E_M, W_M are respectively the

[13] We study heterogeneous waiting costs in Online Appendix V. We consider two scenarios there with observable (general) waiting cost functions and unobservable linear waiting cost functions.

[14] We define social welfare as a simple average of the agents' payoffs for expositional clarity. Our main results hold identically when welfare is defined as a weighted average of the agents' payoffs. See Online Appendix V for details. When welfare is defined by a simple average, then the planner's objective equals a weighted sum of allocative efficiency and total waiting costs, where the weight on the former (latter) summand is positive (negative). This reinterpretation of the objective does not change the main results.

efficiency and welfare of M. Let $\mathsf{M} \subseteq \mathsf{R}^2_+$ be the set of points generated by all individually rational and incentive-compatible mechanisms. The *efficiency-welfare Pareto frontier* is a subset $P \subseteq \mathsf{M}$ of points (e, w) for which there is no $(e', w') \in \mathsf{M}$ such that $e' \geq e$ and $w' \geq w$, with one of the inequalities holding strictly. A point $(e, w) \in P$ is an *extreme point* of the Pareto frontier if (e, w) cannot be written as a convex combination of two other points on the Pareto frontier.

Throughout we impose the following assumptions on the the utility function.

Assumption 2.1. *The utility function u satisfies the following properties:*

i. Spence-Mirrlees property: u is strictly supermodular.

ii. Smoothness: $u(\theta, \omega)$ is absolutely continuous in each argument and its partial derivative u_1 is bounded.

iii. Convexity: The partial derivative $u_1(\theta, \omega)$ is convex in its second argument ω.

iv. $u(0, \omega) = 0$ for all $\omega \in [\omega_0, \omega_n]$.[15]

The first property asserts that agents of higher types appreciate quality more. The second one is a smoothness property that we use in the application of an envelope theorem (Milgrom and Segal, 2002) to characterize allocation rules that are implementable in dominant strategies. The third property is a technical assumption which ensures that the reduced-form version of the planner's problem has an *extreme point* solution. (Details are discussed in the proof approach in Section 3.) The last property is that the lowest type agent $\theta = 0$ derives no utility from being assigned any object. We next review some examples of utility functions satisfying these properties.

Example 2.1 (Multiplicatively separable utility functions)**.** *A commonly considered case satisfying Assumption 2.1 is the case of $u(\theta, \omega) = \theta\omega$, where the agent is an expected utility maximizer. Multiplicatively separable utility functions are ubiquitous in the mechanism design literature.*

Much of the progress in mechanism design is made under the assumption of multiplicative separability (Myerson, 1981; Hartline and Roughgarden, 2008; Condorelli, 2012). Supermodularity of the utility function in our model, however, allows us to capture other scenarios that cannot be easily captured otherwise. For example, consider scenarios where an agent, after observing the allocation mechanism, can change her type by taking costly actions. An agent waiting for housing may decide to find a roommate, or patients waiting for organs may take costly tests or therapies to change their priority on the waitlist (OPTN, 2018). In the next example we demonstrate that our model can capture settings with such endogenous costly actions. This is done by using the supermodularity of the utility functions in our main setup.

Example 2.2 (Utility functions that capture endogenous costly actions)**.** *Consider the case of separable utilities above, with the difference that an agent of type θ can change her type to $\mathsf{t}(\theta, a)$*

[15] Condition iv is not necessary to prove the results concerning direct-revelation mechanisms (since optimal direct-revelation mechanisms are non-wasteful, as discussed in the proof). However, it is necessary for the results about indirect implementations of optimal mechanisms (Section 4), to ensure uniqueness of equilibrium.

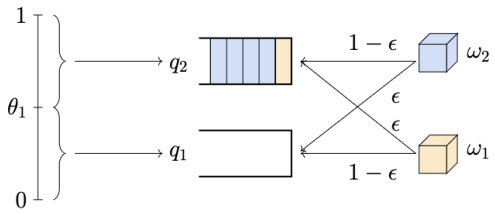


Figure 1: Agent types θ θ_2 join queue q_2, agent types θ_1 $\theta < \theta_2$ join q_1, and agent types $\theta < \theta_1$ take their outside option ω_0.

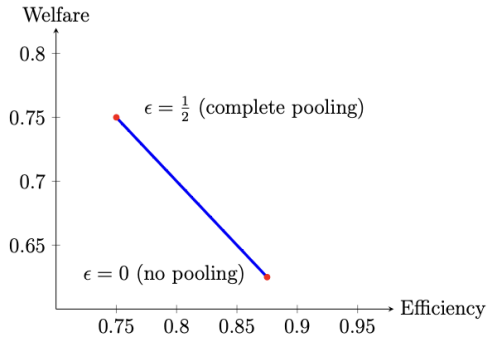


Figure 2: The Pareto frontier corresponding to object types $\omega_2 = 2, \omega_1 = 1$ in Subsection 2.1 is a 45-degree line segment.

by taking an action a 2 A from a compact set of actions A R. The payo of an agent with initial type θ who takes action a, incurs an expected waiting cost of $\bar{c}$, and receives an object with expected quality $\bar{\omega}$ then equals $\mathsf{t}(\theta, a)\,\bar{\omega}$ $\bar{c}$ $\mathsf{c}(a)$*, where the function* $\mathsf{c} : A$! R_+ *gives the cost of taking action a. There is a costless action 0 2 A with* $\mathsf{c}(0) = 0$*. The timing is as follows: agents observe their types, the planner commits to a mechanism, and each agent decides whether to participate in the mechanism and privately chooses an action, which is observable neither by the other agents nor by the designer. Each agent sends a message to the mechanism, and nally the mechanism assigns an allocation timeline to each agent. Online Appendix IV.5 shows that when the function* t *is, e.g., additively separable* $\mathsf{t}(\theta, a) = \theta + a$*, this problem reduces to a problem in our main setup without endogenous costly actions where the utility function u satis es Assumption 2.1 and is not multiplicatively separable. The same holds when* $\mathsf{t}(\theta, a) = \theta a$ *is multiplicatively separable and the cost* $\mathsf{c}(a)$ *is proportional to a^b/b for $b > 1$. This exercise essentially translates some of the ndings of Gershkov et al. (2021) about auctions with endogenous costly actions to our setting.*

2.1 Motivating example: an efficiency-welfare tradeoff

We start with a simple example to demonstrate the economic forces in play. There are two object types $\Omega_+ = \{\omega_1, \omega_2\}$ with equal arrival rates of $\frac{1}{2}$. Thus, $N = 1$. We let $\omega_i = i$. Also, G is the uniform distribution over the unit interval $\Theta = [0, 1]$. The utility function is $u(\theta, \omega) = \theta\omega$.

We consider a class of queuing policies, which we define informally throughout this example. (The formal definitions and main results appear in the next section.) A policy in this class involves two queues q_1, q_2, as demonstrated in Figure 1. Let i 2 $\{1, 2\}$ and i denote $\{1, 2\}$ fig. Upon arrival, an object of type i is sent to $q_{\ i}$ with rate ϵ, and to q_i with rate 1 ϵ, where ϵ 2 $[0, 0.5]$. When an object is sent to a queue, it is allocated in a FIFO manner to one of the agents in that queue; i.e., it is allocated to the agent with the earliest arrival time in the queue. The agent then leaves immediately. Intuitively, higher values of ϵ correspond to more object types being pooled together.

Agents with type below (above) 0.5 are sent to q_1 (q_2). Moreover, the mechanism immediately matches the agents who join q_1 to an object. This is possible because the arrival rates of agents and objects are equal in q_1. What should be the waiting times in q_2 so that this mechanism is incentive compatible? As objects are allocated in a FIFO manner within each queue, all agents who join the same queue incur the same waiting time. Let t_2 be the waiting time in q_2, and define $c_2 = c(t_2)$ as the waiting cost in q_2.[16] For the mechanism to be incentive compatible, c_2 should be high enough so that no agent sent to q_1 would prefer to join q_2, and low enough so that no agent sent to q_2 would prefer to join q_1. By the supermodularity of u, this holds if the agent with type 0.5 attains the same payoff from joining either queue. The *indi erent condition* for this agent is

$$0.5\left((1 \quad \epsilon)\omega_1 + \epsilon\omega_2\right) = 0.5\left((1 \quad \epsilon)\omega_2 + \epsilon\omega_1\right) \quad c_2.$$

Solving this equation and plugging in $\omega_1 = 1, \omega_2 = 2$ gives $c_2 = 0.5 \quad \epsilon$. This equality reveals the tension between efficiency and welfare. Efficiency is maximized by a positive assortative assignment that matches high-type agents to high-type objects. Thus, increasing ϵ reduces efficiency. On the other hand, increasing ϵ reduces the waiting cost c_2 in q_2 (without changing the waiting cost $c_1 = 0$ in q_1), and thus can potentially increase welfare. To investigate this, we compute welfare as $0.25\left((1 \quad \epsilon)\omega_1 + \epsilon\omega_2\right) + 0.75\left((1 \quad \epsilon)\omega_2 + \epsilon\omega_1\right) \quad c_2/2$. This holds because the average agent who joins q_1 has type 0.25, and the average agent who joins q_2 has type 0.75. Plugging in $\omega_1 = 1, \omega_2 = 2$ in the latter equation gives welfare equal to $(5 + 2\epsilon)/8$. Thus, pooling more objects by increasing ϵ also increases welfare, but decreases efficiency. Figure 2 plots the entire range of welfare and efficiency that is attainable by varying ϵ.

What is the structure of optimal mechanisms given arbitrary Pareto weights $\lambda_E, \lambda_W \quad 0$ on efficiency and welfare? To answer this questions it is helpful to look at the efficiency-welfare Pareto frontier. It turns out that every point on the Pareto frontier is implementable by a particular choice of ϵ. (The complete derivation of the Pareto frontier is in Online Appendix IV.1.) The point with the highest efficiency corresponds to $\epsilon = 0$ (no pooling), and the point with highest welfare corresponds to $\epsilon = \frac{1}{2}$ (complete pooling). The optimal solution to the planner's problem is complete pooling if the planner places more weight on welfare $\lambda_W > \lambda_E$, and is no pooling otherwise.

What do optimal mechanisms look like in general?[17] Can they be implemented in practically simple ways? How does their structure change along the Pareto frontier (i.e., as the planner's Pareto weight on welfare increases)? We address these questions in the next three sections.

[16] We note that the *length* of q_2 is the agents' arrival rate to q_2 multiplied by their waiting time; i.e., $\frac{1}{2}t_2$.

[17] Optimality of complete or no pooling does not hold in general. Online Appendix IV.3 gives an example.

3 Characterization of the optimal mechanism

To characterize optimal mechanisms generally in our setup, we will define the classes of *disjoint queue* mechanisms, and then the class of *monotone* disjoint queue mechanisms, which contains the optimal mechanism that we identify. Intuitively, disjoint queue mechanisms use several queues within which objects are allocated to agents in a FIFO manner, as in the motivating example.

Disjoint queue mechanisms. A *disjoint queue* mechanism is characterized by a number k and a strictly increasing sequence of positive reals $\theta_0 = 0, \theta_1, \ldots, \theta_k, \theta_{k+1} = \overline{\theta}$. The mechanism features a finite number of queues $q_0, q_1, \ldots, q_k$. An agent who reports a type θ 2 $[\theta_i, \theta_{i+1})$ for i 0 is assigned upon arrival to the queue q_i. Every object immediately upon arrival is sent to one of the queues $q_0, \ldots, q_k$ to be offered to the earliest-arriving agent in that queue. As objects are allocated to agents in a FIFO manner within each queue, then all agents joining the same queue have the same waiting time. Moreover, (i) for every object type ω and every queue q_i, objects of type ω are sent to q_i at a constant rate (possibly 0) over time, (ii) only objects of type ω_0 are sent to q_0 and every agent assigned to q_0 is immediately allocated, and (iii) the arrival rate of objects to the queue q_i is $G(\theta_{i+1})$ $G(\theta_i)$ for all i 1.

We say that the above mechanism *involves* k *queues*, as q_0 corresponds to agents who take their outside option. In a disjoint queue mechanism, the pool remains at a steady state because the arrival rate of agents to each queue equals the arrival rate of objects to that queue (and agents cannot decline the objects that are assigned to them). Therefore, the length of each queue, as well as the time that a newly arriving agent at that queue has to wait to receive an object, remains constant across time.

Monotone disjoint queue mechanisms. Consider a disjoint queue mechanism. We say that queue q_j has a *higher* type than q_l if $j > l$. Such a mechanism is a *monotone* disjoint queue mechanism if (i) for any object type ω 2 Ω there is at most one queue to which only objects of type ω are sent, and (ii) for any two objects of distinct types, either the two objects are sent to the same queue or the higher-type object is sent to a higher-type queue.[18]

Intuitively, this means that the object types can be partitioned into intervals–which may overlap only at their endpoints–such that object types belonging to the same interval are sent to the same queue. Figure 3 illustrates the assignment of objects and agents to queues in a monotone disjoint mechanism, in the left and right panels, respectively. The two panels demonstrate the same example; hence, the colored areas in the right panel correspond to the object types in the left panel. In a monotone disjoint queue mechanism, the object type ω_0 may be pooled with other object types. (That may happen only in q_1.) This is not the case in the mechanism of Figure 3,

[18] Notably, the class of incentive-compatible monotone disjoint queue mechanisms is strictly smaller than the class of incentive-compatible disjoint queue mechanisms, which is strictly smaller than the class of incentive-compatible mechanisms. We provide examples to demonstrate this in Online Appendix IV.4.

but can happen when the mechanism is changed so that ω_0 is also sent to q_1.[19]

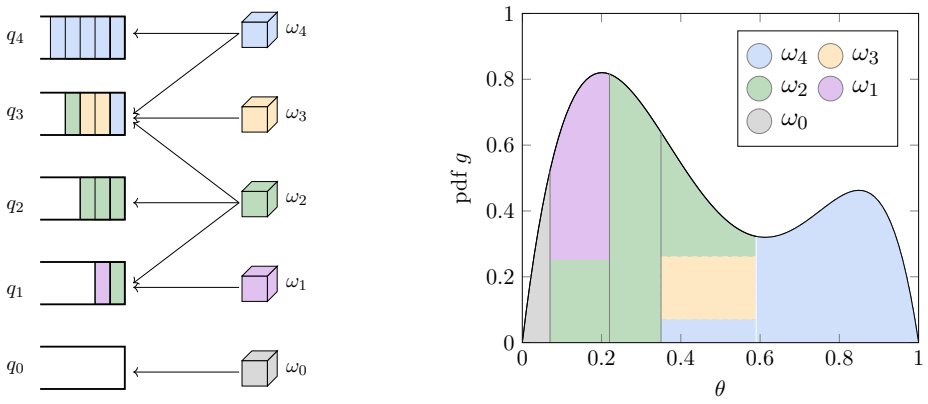


Figure 3: **Left**: An example of a monotone disjoint queue mechanism. Agents joining the same queue are allocated the same expected object type. Higher-type agents join weakly higher-type queues, and higher-type objects are sent to weakly higher-type queues. **Right**: Agents with a type above 0.6 are assigned to the queue q_4 and receive objects of type ω_4 (blue). This threshold, together with three other thresholds, partitions the area below the PDF into five regions. Each region corresponds to a queue in the left figure. Note that agents belonging to the same interval receive the same expected object type and incur the same waiting cost.

Theorem 3.1. *There is an optimal mechanism M that is a monotone disjoint queue mechanism. Moreover, the number of queues involved in M is at most twice the number of object types.*

All proofs are in the appendices. We next discuss the uniqueness properties of monotone disjoint queue mechanisms, distinguishing between the allocation of objects across and within queues.

Allocation of objects across queues. Are there optimal mechanisms which pool objects inconsistently with monotone disjoint queue mechanisms? We show that the answer is negative except for *degenerate* cases where there are a continuum of optimal solutions to the planner's problem. More precisely, in Online Appendix I we show that mechanisms that generate the extreme points of the efficiency-welfare Pareto frontier must pool object types identical with monotone disjoint queue mechanisms. In particular, this implies that any disjoint queue mechanism that generates an extreme point of the Pareto frontier must be a *monotone* disjoint queue mechanism. (Notably, for any non-extreme point of the Pareto frontier, there is an extreme point at which the planner

[19] In that case, some of the agents in q_1 may be assigned ω_0 after waiting. The mechanism can be implemented in an alternative way so that this never happens. The alternative implementation is as follows. Suppose that objects $\omega_0, \ldots, \omega_j$ are pooled together in q_1. Then, when an agent arrives at q_1, the designer immediately determines whether the agent would receive ω_0 or not: ω_0 is assigned to the agent with a fixed probability $p_1 > 0$, and with probability $1 - p_1$ the agent is sent to join q_1. If the agent joins q_1, she receives one of the object types $\omega_1, \ldots, \omega_j$ in a FIFO manner. (The arrival rates of these objects to q_1 remain the same as in the original mechanism.) The parameter p_1 is set so that the rate of agents who join q_1 equals the aggregate arrival rate of objects of types $\omega_1, \ldots, \omega_j$ to q_1. Thus, the expected object type received by the agent is the same as in the original mechanism. Also, the waiting time in q_1 is set so that the agent's expected waiting cost is the same as in the original mechanism.

attains a weakly higher objective value.) The motivating example illustrates this. There are two extreme points on the Pareto frontier in Figure 2. The point with the lowest (highest) efficiency uniquely corresponds to the mechanism that completely pools (does not pool) object types. The other points on the Pareto frontier are not extreme points and can be generated by mechanisms that pool object types inconsistently with monotone disjoint queue mechanisms; i.e., when $\epsilon \in (0, \frac{1}{2})$.

Allocation of objects within each queue. In Online Appendix I, we also establish a uniqueness property for the FIFO queuing scheme used in monotone disjoint queue mechanisms. Loosely speaking, a queuing scheme refers to a probability distribution over the queue positions according to which an arriving object is allocated. We also allow the probability distribution to be *scaled* proportionally to the queue length, such as the service-in-random-order (SIRO) scheme which allocates every object to an agent drawn uniformly at random from the queue. We show that FIFO is the unique scheme that is *universally optimal*. Universal optimality means that, in *any* market, an optimal mechanism that uses schemes other than FIFO can instead adopt FIFO while preserving optimality. A policy such as SIRO is thus not universally optimal. To see why this is the case for SIRO, consider a convex cost function c. Observe that, then, waiting becomes increasingly more costly over time. Incentives for waiting diminish, and some agents may prefer leaving the queue unmatched over waiting longer to be matched. Thus, SIRO does not generally satisfy individual rationality (which is a necessary condition for optimality).

Proof approach for Theorem 3.1. The proof of Theorem 3.1 is based on the notion of *interim* allocation rule $X(\theta)$, which gives the expected object type assigned to an agent with reported type θ. Using standard techniques we write the planner's objective as a functional of X,

$$\int_0^{\bar{\theta}} \Big(\phi(\theta, X(\theta))\lambda_W + u(\theta, X(\theta))\lambda_E \Big) \, \mathrm{d}G(\theta), \tag{3.1}$$

where $\phi(\theta, x) = \theta - \frac{1 - G(\theta)}{g(\theta)} u_1(\theta, x)$. The first summand inside the integral accounts for welfare and the second summand accounts for efficiency. We then consider the *reduced-form* version of the planner's problem: choosing an interim allocation rule X from the set $\mathcal{X}$ of the interim allocation rules of all individually rational and incentive-compatible mechanisms. Classical ironing approaches are not applicable here due to the concurrence of supermodular utility functions and capacity constraints.[20] We take a different approach. It follows from Assumption 2.1 that the planner's objective is a convex functional of X. Then, since $\mathcal{X}$ is compact and convex, the optimal interim allocation rule is an

[20] The reason is that the function $\phi(\theta, x)$ in our problem depends on both the agent type θ and the expected object type x. This is unlike the standard settings (Myerson, 1981; Hartline and Roughgarden, 2008; Condorelli, 2012) where the function counterpart to ϕ depends only on θ, due to multiplicative separability. Toikka (2011) developed a method of two-dimensional ironing for continuously differentiable functionals $J(\theta, x)$ that are weakly concave in the allocation x. However, his formulation does not allow for any form of capacity constraints, due to the pointwise maximization argument used there. This method is thus not applicable to our setup, where objects have finite capacities.

extreme point of X, by Bauer's Maximum Principle (Bauer, 1958).[21] The key step is using a result of Kleiner et al. (2021) that characterizes the structure of extreme points of X, and showing that these extreme points are implementable by monotone disjoint queue mechanisms.

The intuition for the last step is as follows. Let X_{PAM} be the interim allocation rule corresponding to the *positive assortative* assignment, in which objects are allocated to agents so that for any two agents, the higher-type agent is assigned a (weakly) higher-type object. By the result of Kleiner et al. (2021), any extreme point X of X corresponds to a family of disjoint intervals such that X equals X_{PAM} outside the intervals, and inside each interval X equals the average of X_{PAM} over that interval. From this structure, it follows that every extreme point of X can be implemented by a monotone disjoint queue mechanism: the mechanism pools inside each interval but not outside the intervals. Appendix A contains a more detailed proof outline and the formal proof.[22]

4 Implementation through information design

We next provide an alternative implementation of the optimal mechanism. As in common waitlist mechanisms in practice, this implementation uses a single queue and allows agents to decline an assignment. We show that such an implementation is possible if the planner can control the information disclosed to agents about objects. Then, the optimal mechanism can be implemented by a first-come first-served (FCFS) waitlist with deferrals, together with an information disclosure policy for objects that *pools adjacent object types*.

Before formally presenting this result, we present an overview. In a FCFS waitlist with deferrals, every agent enters the same queue and waits for an object to be offered to her. If the agent accepts an offer, she departs the market immediately; if she declines an offer, she waits for the next offer. Every object is offered to agents in order of their arrival and is assigned to the agent with the earliest arrival time who accepts the offer.

Moreover, the planner can decide what information to disclose to agents about an object. One possible disclosure policy is, e.g., providing to all agents not the exact object type, but rather an interval $[\omega_l, \omega_r]$ that contains the object type. An information disclosure policy that *pools adjacent object types* partitions the space of object types $[\omega_0, \omega_n]$ to a number of such intervals.

We find that the optimal monotone disjoint queue mechanism M can be implemented by a FCFS waitlist with deferrals paired with an information disclosure policy that pools adjacent object types. The interim payoff of every agent type under this implementation equals the interim payoff of the same agent type in M. Proving this result requires analyzing a dynamic game played among the agents: in the FCFS waitlist, an agent's decision about whether to accept an offered

[21] By Bauer's Maximum Principle, an upper semicontinuous convex real functional h over a compact convex subset K attains its maximum at an extreme point of K.

[22] While this proof approach is not constructive, constructing optimal mechanisms are possible when the utility function is multiplicatively separable. This is demonstrated in Online Appendix II.

object depends on how the agents ahead of her make decisions. The proof establishes that in every equilibrium of this game, any agent type attains the same payoff as in M . We first provide intuition for this and then present the formal results. For intuition, below we consider the agents' incentive constraints in monotone disjoint queue mechanisms.

Fact 4.1. *An incentive-compatible monotone disjoint queue mechanism is* dynamically incentive-compatible*, in the following sense: Let t_i be the waiting time in queue i. Consider an arbitrary agent of type θ who joins queue i. Then, for every queue j and time t $\min\{t_i, t_j\}$, it holds that*

$$u(\theta, \overline{\omega}_i) \quad (c(t_i) \quad c(t)) \quad u(\theta, \overline{\omega}_j) \quad (c(t_j) \quad c(t)).$$

Intuitively, dynamic incentive compatibility means that an agent, at any point after arrival, prefers to remain in the queue that she has joined initially, rather than switching to a different queue while preserving her current waiting time. The above inequality is derived by subtracting $c(t)$ from both sides of the agent's incentive constraint upon arrival, which states that the agent's expected payoff from joining queue i is no less than her payoff from joining queue j. Notably, the dynamic incentive compatibility of monotone disjoint queue mechanisms relies on assigning objects to agents within each queue in a FIFO manner. For example, if an arriving object is assigned to a waiting agent in a SIRO manner, then the resulting mechanism is generally not dynamically incentive compatible or individually rational. (This is shown in Online Appendix IV.2.)

From the above observation, we can intuit why the optimal monotone disjoint queue mechanism M can be implemented by a FCFS waitlist with deferrals. Consider the disclosure policy that partitions the space of object types into intervals in the same way as in M , and discloses only the interval to which an arriving object's type belongs. At any given time, each agent on the FCFS waitlist chooses an *acceptance* threshold and accepts only offers with expected type above her time-specific threshold. Since the incentive compatibility constraint holds dynamically in M , it follows that there is a steady-state equilibrium for the waitlist where an agent's acceptance threshold remains constant across time and equal to the expected object type that she receives in M . The bulk of the analysis proves the uniqueness of this equilibrium.

We next state the formal definitions for information disclosure policies and steady-state equilibrium, followed by the main finding of this section.

4.1 Information disclosure policies

An *information disclosure policy* μ consists of a finite signal realization space $\widehat{\Omega}$ and a family of distributions $\{\mu(\cdot|\omega)\}_{\omega \in \Omega}$ over $\widehat{\Omega}$. For brevity, we sometimes call μ a disclosure policy and $\widehat{\Omega}$ a realization space. When an object of type ω arrives, the planner draws a realization from $\mu(\cdot|\omega)$, and discloses only that realization to all agents.

Let $\Omega_\mu \quad \mathbb{R}_+$ be the set that contains, for every $s \in \widehat{\Omega}$, the expected object type conditional on

receiving s. The disclosure policy μ is *nondegenerate* if every element of $\widehat{\Omega}$ is realized with positive probability and $\widehat{\Omega}$ and Ω_μ are the same size; i.e., there are no distinct signal realizations $s_1, s_2 \in \widehat{\Omega}$ such that the expected object type conditional on receiving s_1 equals the expected object type conditional on receiving s_2.

Definition 4.1. *We say* μ pools adjacent object types *if it is nondegenerate and the realization space of* μ *is a finite family of intervals* $\widehat{\Omega} = \{[\delta_{2i}, \delta_{2i+1}]\}_{i=0}^{k}$ *for* $k \geq 0$*, such that*

(i) $\delta_j \in \Omega$ *and* $\delta_j < \delta_{j+1}$ *for all* j*;*

(ii) for all $\omega \in \Omega$ *there exists* i *such that* $w \in [\delta_{2i}, \delta_{2i+1}]$*, and*

(iii) $[\delta_{2j}, \delta_{2j+1}] \in \widehat{\Omega}$ *is realized only if the arriving object's type belongs to this interval.*

Intuitively, such a disclosure policy uses a set of intervals to partition $[\omega_0, \omega_n]$. These intervals are disjoint except possibly at their endpoints. If an object type is an interior point of such an interval $[\delta_{2i}, \delta_{2i+1}]$, then the policy sends the signal $[\delta_{2i}, \delta_{2i+1}]$. Otherwise, the object type must be the endpoint of at most two intervals. In that case, the policy discloses one of the two intervals.

4.2 The FCFS waitlist with deferrals under an information disclosure policy

We describe how the planner runs a FCFS waitlist with deferrals under a disclosure policy μ. Intuitively, for an arriving object of type $\omega \in \Omega$ a single realization is drawn from $\mu(\cdot|\omega)$, which is revealed to any agent to whom the object is offered. The object is offered to agents in the order of their arrival time and is assigned to the agent who accepts it first. We model this process as a dynamic game, formally define its steady-state equilibrium, and show its uniqueness.

We refer to an agent with waiting time t as an agent at *position* t (resembling positions on a waitlist). We let $T = \mathbb{R}_+$ denote the set of all positions. Upon receiving an offer, the agent computes the expected type of the offered object conditional on the received signal realization and decides whether to accept it. This conditional expectation is also called the *interim type* of the object. Thus, Ω_μ is the set of all objects' interim types. The rate at which objects with interim type $x \in \Omega_\mu$ arrive is $N \sum_{\omega \in \Omega} F(\omega)\mu(x|\omega)$, which we denote by $F_\mu(x)$.

We restrict attention to steady-state equilibria where the agents make decisions deterministically and where agents of the same type make the same decisions: An agent of type θ has a *decision rule* $D_\theta : \mathbb{R}_+ \to \mathbb{R}_+$ such that $D_\theta(t)$ is the lowest interim object type that she accepts when she is at position t. This is also called the agent's *acceptance threshold* at t. The possible dependence of the agent's decision rule on the state of the waitlist is suppressed from the notation, as we study the steady-state equilibria where the state of the waitlist does not change over time. Without loss of generality, we assume that the range of D_θ belongs to Ω_μ. As we will see shortly, every agent chooses her decision rule in order to maximize her expected payoff.

Definition. A *steady-state equilibrium of the FCFS policy under a disclosure policy* μ is defined by a tuple $(A, b, m, \bar{t}, \{D_\theta\}_{\theta \in \Theta})$ that satisfies the following properties:

(i) **The densities of agents at all positions are well-defined:** For each $t \in T$, $A(t)$ is a measure defined on the Lebesgue measurable subsets of Θ and has a density function $a(t, \cdot) : \Theta \to \mathbb{R}_+$. The function $a(t, \theta)$ gives the density of agents of type θ with waiting time t. Furthermore, for each $\theta \in \Theta$, $a(\cdot, \theta) : T \to \mathbb{R}_+$ is left-continuous, and its support is in $[0, \bar{t}]$, where $\bar{t} \in \mathbb{R}_+$. The finiteness of $\bar{t}$ ensures that agents wait for a bounded amount of time.

(ii) **The densities of objects offered to agents at every position are well-defined:** $b : T \times \Omega_\mu \to \mathbb{R}_+$, and for all $x \in \Omega_\mu$, $b(\cdot, x)$ is right-continuous and increasing. We denote by $b(t, x)$ the measure of objects with an interim type x that are offered to agents with waiting time t.[23] It holds that $b(t, x) = F_\mu(x)$ for $t \geq \bar{t}$; i.e., the rate at which objects of interim type x are offered to the waitlist equals the arrival rate of such objects.

(iii) **The allocation rates of objects are well-defined and feasible:** m is a measure defined on the Lebesgue measurable subsets of $\mathbb{R}^3$ and assigns a nonzero measure to a subset only if it is inside $T \times \Omega_\mu \times \Theta$. (The interpretation is that $\mathsf{m}(T', \Omega', \Theta')$ is the measure of objects of a type in $\Omega' \subseteq \Omega$ that are matched to agents of a type in $\Theta' \subseteq \Theta$ at positions $T' \subseteq T$.) The marginal measure induced by m over Θ is the distribution of agent types G.[24] Also, the marginal measure induced by m on $x \in \Omega_\mu$ is at most the arrival rate of x; i.e., $\mathsf{m}(T \times \{x\} \times \Theta) \leq F_\mu(x)$. For all $t \in T$, $\mathsf{m}(\{t\} \times \{x\} \times \Theta) \leq b(t, x)$; i.e., the total rate of objects of interim type x allocated to agents at position t is no more than the rate such objects are offered at position t. Moreover, the restriction of m to $\{t\} \times \{x\} \times \Theta$ has a density, which we denote by $m(t, x, \cdot)$.[25]

(iv) **The agents' decisions are respected:** $D_\theta : T \to \Omega_\mu$ is left-continuous for all θ and $m(t, x, \theta) > 0$ only if $D_\theta(t) \geq x$.

(v) The *balance equations*, the *market clearing* condition, and *optimality of agents' decisions* hold, which are defined as follows.

Balance equations. The balance equation for agents states that for any $t \in T$ and $\theta \in \Theta$, $a(t, \theta) = \sum_{x \in \Omega_\mu} m(t, x, \theta) = \lim_{t' \uparrow t^+} a(t', \theta)$. The interpretation is that the agents depart only when matched. Let $M(t, x)$ denote $m(\{t\}, \{x\}, \Theta)$, the rate at which objects with interim type x are assigned to agents at position t. The balance equation for objects states that for any $t \in T$ and $x \in \Omega_\mu$, $b(t, x) = M(t, x) = \lim_{t' \uparrow t} b(t', x)$. The interpretation is that objects remain available until they are allocated to an agent.

Market clearing. To state this condition we first define the following notions. An agent type θ *demands* an interim object type x at position t if $D_\theta(t) \geq x$ and $a(t, \theta) > 0$. When this condition

[23] The function $b(\cdot, x)$ being increasing means that there are fewer such objects left to offer to agents with shorter waiting times.

[24] That is, for every Lebesgue measurable $\Theta' \subseteq \Theta$, $\mathsf{m}(T, \Omega_\mu, \Theta')$ equals the measure assigned by G to Θ'.

[25] That is, $\mathsf{m}(t, x, \Theta') = \int_{\theta \in \Theta'} m(t, x, \theta) \mathrm{d}\theta$ for all measurable $\Theta' \subseteq \Theta$.

holds for some $\theta \in \Theta$, we say x is *demanded* at position t. We say x is *supplied* at position t if $b(t,x) > 0$. The *market clearing* condition states that

(i) *the set of agents demanding an object is well-defined:* For every position t and interim object type x, the set of agent types who demand x at t is Lebesgue measurable;

(ii) *agents who demand objects with excess supply are matched:* If $\int_{\theta \in \Theta} m(t,\theta,x)\,\mathrm{d}\theta < b(t,x)$ for $x \in \Omega_\mu$, then for every agent type θ demanding x at position t, $\sum_{y \in \Omega_\mu} m(t,\theta,y) = a(t,\theta)$;

(iii) *the most-demanded object type at each position is fully allocated:* If a positive $x \in \Omega_\mu$ is the highest interim object type demanded at position t, then $\int_{\theta \in \Theta} m(t,\theta,x)\,\mathrm{d}\theta = b(t,x)$; and

(iv) *at each position, the allocation of objects to agents is agnostic to agent types:* For all positive $x \in \Omega_\mu$, and agent types θ, θ' who demand x at position t, $\frac{m(t,\theta,x)}{a(t,\theta)} = \frac{m(t,\theta',x)}{a(t,\theta')}$.

We denote the latter quantity by $P_t(x)$ when x is allocated to some agent type at position t. Then, this quantity gives the probability that an agent allocates x, conditional on being at position t and demanding x at that position. To define the optimality of agents' decisions next, it helps to define $P_t(x)$ for every other x as well: For the largest $x \in \Omega_\mu$ that is supplied at position t but not allocated to any agent type, $P_t(x) = 1 \quad \sum_y P_t(y)$, where the summation is over all $y \in \widehat{\Omega}_\mu$ that are allocated to some agent type at position t. For all other x, $P_t(x) = 0$.

Optimality of an agent's decision. Recall that the decision rule of an agent is a left-continuous function from T to Ω_μ. Let D be the set of all such functions. The payoff of an agent with type θ under any decision rule $D \in \mathsf{D}$ is $\pi_\theta(D) = u(\theta, \omega_D) \quad t_D$, where ω_D, t_D respectively denote the expected object type allocated to the agent and her expected waiting cost under decision rule D. (The expectations ω_D, t_D are computed according to the conditional probabilities $P_t(x)$ that give the probability that an agent demanding x at position t receives x at that position.) At a steady-state equilibrium, every agent of type θ chooses a decision rule $D_\theta \in D$ that maximizes her payoff. This completes the definition of a steady-state equilibrium.[26]

Fixing a disclosure policy μ, we say that the steady-state equilibrium is *essentially unique* under μ if there is a finite subset of agent types $\Theta' \quad \Theta$ for which the following holds: for every $\theta \in \Theta \cap \Theta'$ there exist $a_\theta, b_\theta \in \mathsf{R}_+$ such that the expected object type allocated to any agent of type θ in any steady-state equilibrium is a_θ, and the expected waiting cost incurred by any such agent is b_θ. We note that the expectations are computed at the agent's arrival time.

Proposition 4.1. *For any disclosure policy μ, the FCFS waitlist with deferrals under μ has an essentially unique steady-state equilibrium. Moreover, for any two agents at any steady-state equilibrium, the agent with the higher type is allocated an object with a weakly higher interim type.*

[26] In the analysis (Appendix B) we show that, despite the apparently wide range of possibilities for the choice of m, this measure takes a trivial form at a steady-state equilibrium: at every position $t > 0$, there is at most one $\omega^t \in \Omega_\mu$ that is both demanded and supplied. Furthermore, the demand for ω^t at position t equals the supply of it, i.e., the measure assigned by $A(t)$ to agent types demanding ω^t is $b(t,\omega^t)$.

A *full-disclosure* policy discloses a distinct signal realization for every object type. A *no-disclosure* policy discloses the same signal realization for all object types in Ω_+.[27] By Proposition 4.1, there is an essentially unique steady-state equilibrium under a no-disclosure policy where all agents who wait, wait the same amount of time, and are allocated the same object type in expectation. Moreover, the equilibrium under the full-disclosure policy is *positive assortative* and therefore maximizes allocative efficiency, but can be suboptimal otherwise. There are markets where neither a full-disclosure nor a no-disclosure policy is optimal (Example IV.1 in the appendix). Nevertheless, the FCFS waitlist with deferrals can attain the optimal solution to the planner's problem under an appropriate information disclosure policy, as shown below.

4.3 The structure of optimal disclosure policies

By Proposition 4.1, all steady-state equilibria of the FCFS waitlist with deferrals under μ generate the same level of welfare (i.e., the average of agents payoffs), and also the same level of efficiency (i.e., the average of agents' utilities). Thus, all such steady-state equilibria generate the same objective value for the planner. If this objective value equals the objective value attained by an optimal mechanism, then we say μ *attains the second-best objective value*.

Theorem 4.1 (Indirect Implementation of the Optimal Mechanism)**.** *There is an information disclosure policy μ that pools adjacent object types and attains the second-best objective value.*

Proving the theorem requires analyzing the dynamic game played among the agents in the FCFS waitlist. The proof more generally shows that, for any disjoint queue mechanism M, there is a disclosure policy μ_M that attains the same objective value as M. Such a disclosure policy would pool object types in the same way as in M. A main step in the proof is showing the existence and uniqueness of the steady-state equilibrium of the game played among the agents (Proposition 4.1). The intuition for existence was discussed earlier: a dynamic version of incentive constraints holds in disjoint queue mechanisms (Fact 4.1). To prove uniqueness, we first show that any steady-state equilibrium is *positive assortative with respect to* μ_M, in the sense that the interim type of an object received by a higher-type agent is no smaller than that obtained by a lower-type agent. This is due to the agents' incentive constraints across time. Then, we show that the agents' incentive constraints uniquely pin down a waiting time for obtaining each interim object type.

5 The efficiency-welfare Pareto frontier and comparative statics

We next study how the structure of optimal solutions to the planner's problem changes along the efficiency-welfare Pareto frontier. We focus on multiplicatively separable utility functions throughout this section. The high-level finding is a monotone comparative statics result: a higher weight

[27] Thus, the realization space of a no-disclosure policy μ has size at most 2, and contains s with $\mu(s|\omega) = 1$ for $\omega \in \Omega_+$.

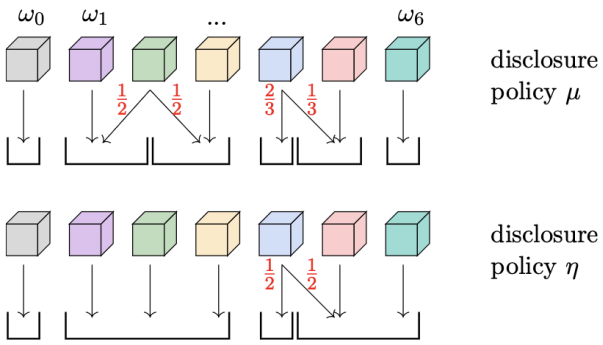


Figure 4: The signal realizations in μ and η are demonstrated by the brackets underneath the objects. They have an interval structure. The probabilities with which a signal realization is sent conditional on the realized object type are indicated. Unmarked arrows indicate a probability of 1. For example, there is a unique signal realization in μ that is sent with probability 1 if the object type is ω_6. The disclosure policy η is less informative than the disclosure policy μ, and the intervals that pool (at least two) object types in η are coarser than the ones in μ.

on welfare in the planner's objective leads to *pooling more* objects in the optimal solution. In the context of information disclosure policies, this means the intervals that pool object types become *coarser* when the weight on welfare increases. This results in a less informative disclosure policy in the sense of Blackwell (1953).[28] After presenting this result, we study the relation between shortage of supply and informativeness of disclosure policies.

5.1 Informativeness of the solutions along the Pareto frontier

Given a disclosure policy μ, every signal realization induces a posterior belief over the distribution of the object types. In our setup, an agent's utility depends on posterior beliefs only through the posterior mean. Recall from Section 4.2 that we call this posterior mean the object's interim type. Thus, an information disclosure policy can be identified by the distribution of posterior means (the objects' interim types) that it induces. Denote this distribution by D_μ. We say a disclosure policy μ is *more informative* than η if D_μ is a mean-preserving spread of D_η. This is the canonical ranking proposed by Blackwell (1953).[29]

When both μ and η pool adjacent object types, then the above condition has a more intuitive description: intervals that pool (at least two) object types are *coarser* in η than in μ. Intuitively, each such interval in μ is contained in an interval in η, as demonstrated in Figure 4.[30]

[28] As we will see, the comparative statics result holds for all optimal disclosure policies, even those that do not pool adjacent object types.

[29] Alternatively, one can define μ to be more informative than η if η is a *garbling* of μ (Blackwell, 1953; Gentzkow and Kamenica, 2016). The results in this section remain essentially the same under this alternative definition.

[30] Remark C.1 in the appendix formalizes this intuition.

Theorem 5.1. *Let the utility function be multiplicatively separable, and x the planner's Pareto weight on e ciency. Then, any disclosure policy that attains the second-best objective value for a smaller Pareto weight on welfare is more informative than any policy that attains the second-best objective value for a larger Pareto weight.*

The theorem follows from a result of Nikzad (2023). We include a proof in the appendix for completeness. The theorem shows that there is a tradeoff between welfare and informativeness on the Pareto frontier: implementing outcomes with higher welfare requires less informative disclosure policies. For example, the point with the highest welfare on the Pareto frontier is implemented by the least informative disclosure policy, and the point with the highest efficiency is implemented by the most informative one—full disclosure. The proof further shows that the extreme points of the Pareto frontier (e.g., the points at which the Pareto frontier is strictly concave) can be implemented by an essentially unique disclosure policy, which pools adjacent object types. For the proof, one needs to also consider the non-extreme points of the Pareto frontier, which are implementable by disclosure policies that, in general, do not pool adjacent object types.

The proof is in Appendix C. We first write the reduced form of the planner's problem (3.1) which, due to the separability of the utility function, simplifies to maximizing a linear functional

$$\max_{X \in \mathcal{X}} \int_0^{\bar{\theta}} X(\theta) \left(\lambda_E \theta + \lambda_W \frac{1 \quad G(\theta)}{g(\theta)} \right) \mathrm{d}G(\theta). \tag{5.1}$$

We call the coefficient of $X(\theta)$ the *virtual value* function. Classical ironing approaches apply here: An optimal solution X can be described by a family of disjoint *ironed* intervals, where the agent types inside each interval are pooled (i.e., are allocated the same expected object type by X). Outside the intervals the allocation of objects to agent types is positive assortative. When the weight λ_E on efficiency goes up, the virtual value function becomes *more increasing*, in the sense that it has a pointwise larger derivative. The novel finding is that, then, each ironed interval in the optimal solution shrinks from both sides. In our problem, that translates to a smaller subset of object types being pooled together, and thus a more informative disclosure policy.

5.2 Markets with severe shortage of supply

We next study markets where the supply is severely short. The welfare and efficiency of any mechanism remains the same if the arrival rates of agents and objects are scaled at the same rate. Motivated by this observation, the next result fixes the agents' arrival rates and shows that, when there is a severe shortage of supply, the optimal policy takes a simple form: either a full- or a no-disclosure policy is optimal.

Proposition 5.1. *Let the utility function be multiplicatively separable, and suppose the PDF g has full support and is continuously di erentiable everywhere in its domain. For any $\epsilon > 0$ there is*

a threshold $N_\epsilon > 0$ such that, when the objects' arrival rate is smaller than N_ϵ, a full-disclosure (no-disclosure) policy attains the second-best objective value if $\lambda_W < \lambda_E \quad \epsilon$ ($\lambda_E < \lambda_W \quad \epsilon$).

In any market, the Pareto frontier is sandwiched between two 45-degree lines with slope 1 which lie above and below the Pareto frontier. The above proposition implies that these two lines become identical at the limit $\epsilon \; ! \; 0$. Then, the Pareto frontier is an affine 45-degree line segment, as in the motivating example (Figure 2), and either full disclosure or no disclosure would be optimal. Thus, the proposition can also be formulated slightly differently: restricting attention to these two simple disclosure policies comes with a loss; however, the loss approaches 0 as the agents' arrival rate grows large.

The proposition is proved is in Online Appendix II. The proof intuition is that, when N is sufficiently small, the virtual value function is monotonic over the set of agent types that receive an object. If the function is increasing, then a positive assortative allocation of objects to agents maximizes the planner's objective, and thus full disclosure is optimal. If the function is decreasing, then monotonicity binds in the planner's problem (5.1), and the optimal solution thus pools all agent types that receive an object. That is, a no-disclosure policy is optimal. To determine which of the two policies is optimal based on the Pareto weights, we investigate the derivative of the virtual value function when N approaches 0.

What happens when the supply shortage is not extreme? That depends on the shape of the virtual value function. We elaborate on the answer below, focusing on welfare maximization.

Moderate shortage of supply. We first define two disclosure policies which turn out to be optimal here. An *upper-censorship* policy fully reveals an object's type ω if the type is below a fixed threshold, and otherwise reveals only that the object type is above that threshold. That is, there is $\omega \; 2 \; \Omega$ such that the disclosure policy sends a distinct signal realization s_ω if $\omega < \omega$, and sends a signal realization $\bar{s}$ if $\omega > \omega$. When $\omega = \omega$ the disclosure policy may send either signal s_ω or $\bar{s}$. A *lower-censorship* policy is defined similarly, except that it reveals an object's type if it is above a fixed threshold, and otherwise reveals only that the object type is below the threshold. Let $h(G) = \frac{g(\theta)}{1 \; G(\theta)}$ denote the hazard rate of G.

Proposition 5.2. *Suppose the utility function is multiplicatively separable, and that the planner's objective is maximizing welfare. If h_G is single-peaked (single-dipped)*[31] *then for every value of objects' arrival rates $N \quad 1$ there is a lower-censorship (upper-censorship) disclosure policy $\mu(N)$ that attains the second-best objective value. Moreover, $\mu(N)$ is more (less) informative for lower values of N.*

We remark that the hazard rate of G is single-peaked for commonly known distributions G

[31] A function $z : [0, \bar{\theta}] \; ! \; \mathsf{R}$ is single-peaked (single-dipped) if there is $\theta \; 2 \; \Theta$ such that z is increasing (decreasing) over $[0, \theta \;]$ and decreasing (increasing) over $[\theta \;, \bar{\theta}]$.

such as Beta, Cauchy, Gamma, Log-normal, Normal, and Weibull.[32] The same holds when these distributions are truncated to have a nonnegative support. The welfare-maximizing solution can thus be implemented by a lower-censorship disclosure policy when G is such a distribution, and a higher imbalance then corresponds to a more informative disclosure policy.

The complete proof is in Online Appendix II. The proof notes that the reduced-form problem (5.1) boils down to $\sup_{X \in \mathcal{X}} \int_0^{\bar{\theta}} X(\theta)/h_G(\theta)\, \mathrm{d}G(\theta)$ for a welfare-maximizing planner. If h_G is single-peaked, then the virtual value function is decreasing below a threshold and increasing above it. It follows that in the reduced-form problem monotonicity of X binds for agent types below a threshold θ , but not for those above it. Hence, the agent types below θ are pooled together and receive the same expected object type in the optimal solution, whereas the allocation of objects to agent types above θ is positive assortative; i.e., lower censorship is optimal. Why is this policy more informative for lower values of N? Intuitively, with a greater supply shortage, a smaller fraction of agents with type below θ receive an object. Thus, a smaller fraction of objects are allocated to agents who are pooled together, which implies that the disclosure policy is more informative.

6 Discrete markets

We provide an analogous result for the optimality of monotone disjoint queue mechanisms in discrete markets. A naive adaptation of monotone disjoint queue mechanisms is not optimal or even incentive compatible in the discrete setup. The reason is that, unlike in the steady state of the continuum model, the incentive constraints themselves change dynamically as queue lengths do. Thus, we adjust the mechanism to make it incentive-compatible: rather than assigning each agent to the same queue that she is sent to in the continuum model based on her type, we assign her to the queue that maximizes her expected payoff conditional on the present queue lengths. We prove that the adapted mechanism is optimal in large discrete markets. We also provide simulations for convergence rates in Online Appendix III.

The discrete setup. In the discretized version of our model time starts at 0, and agents and objects arrive at the market over time, according to independent Poisson processes with rates ζ and ζN, respectively. We call ζ the *market size*. The objects' and agents' types are drawn i.i.d. from distributions F and G, respectively. Similar to our main setup, the only unobservable component of the market to the planner is the agents' types: an agent's type is known only to herself.

An arriving agent reports a type to the mechanism, which in turn promises the agent an allocation timeline. This allocation timeline can depend on the entire history of the market from the beginning of time. (Thus, unlike in the continuum setting, agents of the same type may be promised different allocation timelines.) The payoff of an agent is her utility from the expected

[32] In contrast, Hjorth and additive Weibull distributions can have a "bathtub-shaped" hazard rate, which is single-dipped. Such distributions are studied in various contexts (Leithead, 1970; Al Abbasi et al., 2019; Thach, 2022).

object type she receives minus her expected waiting cost. In an *incentive-compatible* mechanism the expected payoff of every participating agent is maximized if she reports her true type. An incentive-compatible mechanism is called *individually rational* if the expected payoff of any participating agent is nonnegative when she reports her true type to the mechanism.

We focus on analyzing *steady-state* mechanisms. In words, a mechanism M in the discrete setup is a steady-state mechanism if, for an agent of type θ who arrives at time t, the expected object type and the expected waiting cost in the promised allocation timeline to that agent converge in distribution to fixed distributions that depend only on θ, as t goes to infinity. The formal definition of this condition, as well as the proofs for the following results, appear in Online Appendix III.

Monotone disjoint queue mechanisms for discrete markets. Similar to the main setup, monotone disjoint queue mechanisms are defined by a finite number k of queues, namely $1, \ldots, k$. Every object of type $\omega \in \Omega_+$ immediately upon arrival is sent to queue i independently with probability $p_{\omega,i}$, where $\sum_{i=1}^{k} p_{\omega,i} = 1$. These probabilities are set so that, for any two objects of distinct types, the higher-type object is never sent to a lower-type queue. The queue lengths at all times and the probabilities $p_{\omega,i}$ are common knowledge.

When an object is sent to a queue, it is assigned to the agent in the queue with the longest waiting time. If there is no agent waiting in that queue, the object is discarded.[33] Upon arrival, an agent is sent to the queue that maximizes her expected payoff, which is computed conditional on the queue lengths at the agent's arrival time. If there are multiple such queues, the agent is sent to the queue with the highest index. If none of the queues has a positive expected payoff for the agent, then she is sent to queue 0, where she immediately receives her outside option ω_0.

We next show that, in large markets, this simple class of mechanisms can achieve the highest objective value among all of the mechanisms belonging to the more complex class of steady-state direct-revelation mechanisms, which can condition on the entire market history. We let the market size ζ grow and hold the rest of the parameters (N, F, G, c) fixed. For a given ζ, let the *second-best objective value*, denoted by $V(\zeta)$, be the supremum of the planner's objective over all steady-state, incentive-compatible, and individually rational mechanisms. Also, let $\tilde{V}(\zeta)$ be the supremum of the planner's objective over all monotone disjoint queue mechanisms.

Proposition 6.1. *With linear waiting costs, $c(t) = \gamma t$ for $\gamma > 0$, monotone disjoint queue mechanisms are optimal in large discrete markets; that is, $\lim_{\zeta \to \infty} \tilde{V}(\zeta) = \lim_{\zeta \to \infty} V(\zeta)$.*

The planner can also implement the optimal mechanism indirectly, without eliciting the agents' types. In the indirect implementation, each agent observes the real-time queue lengths (or just the expected waiting time at each queue upon arrival) and joins the queue that maximizes her expected payoff. This induces the same stochastic process as that induced by the monotone disjoint queue

[33] Note that this assumption only reduces the planners' objective. As we will show, monotone disjoint queue mechanisms are optimal in large markets, and thus the *waste* generated under this assumption vanishes in large markets. We also make this observation in the simulations of Online Appendix III.1 that dismiss this assumption.

mechanism. Intuitively, this process can be seen as a *tâtonnement* process. Fewer (more) agents join a queue when it becomes *too long* (*too short*). The queue lengths are adjusted over time so that, on average, the demand equals supply in each queue. The average queue lengths converge to the (unique) market-clearing queue lengths in the continuum model, when the market becomes sufficiently large.[34]

For the proof, we (i) show that the planner's objective in the continuum setting is a large-market upper bound for the planner's objective in the discrete setting, (ii) adapt the optimal mechanism from the continuum setting to the discrete setting as described earlier, and (iii) show that the adapted mechanism indeed attains the upper bound from the continuum setting using the concentration bounds of Ashlagi et al. (2022a) for the waiting costs. The proof is in Online Appendix III. There, we also investigate the speed of convergence through simulations.

7 Conclusion

This paper considers the allocation of vertically differentiated objects to agents with private information. The optimal mechanism uses a set of FIFO queues (without deferrals), where each queue offers a lottery over adjacent types of objects. When partial disclosure of information about objects is viable, a simple implementation is possible by a single FCFS waitlist with deferrals. Such FCFS or priority-based waitlist mechanisms are ubiquitous, for example, in the allocation of deceased-donor organs. There is a tradeoff between welfare and informativeness: achieving higher welfare requires pooling larger subsets of object types.

When there is observable horizontal differentiation across different groups of agents (but not within groups), objects can still be optimally allocated to each agent group using such FCFS waitlists. It is intriguing to study settings in which preferences exhibit richer horizontal idiosyncrasies.

In our mechanisms, after getting allocated agents are not allowed to renege and rejoin the queue. This could be rationalized by assuming that agents are identifiable. Another limitation is that, although FCFS seems simple, it may be difficult for agents to reason about the beliefs of other agents ahead of them on the waitlist. However, a social planner can disclose average waiting times for different (possibly pooled) types of objects. Another concern is the ability to partially disclose information about objects. This can motivate the study of optimal mechanisms given constraints over information disclosure policies.

[34] Notably, the *tâtonnement* process computes average queue lengths over time without requiring any knowledge about the distribution of agent types. Suppose, e.g., that the distribution of agent types is Normal with known mean but unknown variance. The planner can first use a monotone disjoint queue mechanism with no pooling, and let the process run until average queue lengths converge. Consider the agent type who is indifferent between joining the lowest queue and taking her outside option in the continuum model. By plugging the average queue length for the lowest queue into the agent's indifference condition, one can approximate the unknown mean with an arbitrarily small error. Thus, the planner can first learn the distribution of agent types using a simple monotone disjoint queue mechanism, and then compute and implement the optimal one.

References

Nikhil Agarwal, Itai Ashlagi, Michael A Rees, Paulo Somaini, and Daniel Waldinger. Equilibrium allocations under alternative waitlist designs: Evidence from deceased donor kidneys. *Econometrica*, 89(1):37–76, 2021.

Mohammad Akbarpour, Shengwu Li, and Shayan Oveis Gharan. Thickness and information in dynamic matching markets. *Journal of Political Economy*, 128(3):783–815, 2020.

Jamal N Al Abbasi, Mundher A Khaleel, Moudher Kh Abdal-hammed, Yue Fang Loh, and Gamze Ozel. A new uniform distribution with bathtub-shaped failure rate with simulation and application. *Mathematical Sciences*, 13:105–114, 2019.

Jerry Anunrojwong, Krishnamurthy Iyer, and Vahideh Manshadi. Information design for congested social services: Optimal need-based persuasion. In *21st ACM Conference on Economics and Computation*, 2020.

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

Itai Ashlagi, Jacob Leshno, Pengyu Qian, and Amin Saberi. Price discovery in waiting lists: A connection to stochastic gradient descent. Technical report, working paper, 2022a. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4192003.

Itai Ashlagi, Afshin Nikzad, and Philipp Strack. Matching in dynamic imbalanced markets. *Review of Economic Studies*, 2022b.

Mariagiovanna Baccara, SangMok Lee, and Leeat Yariv. Optimal dynamic matching. *Theoretical Economics*, 15(3):1221–1278, 2020.

Heinz Bauer. Minimalstellen von funktionen und extremalpunkte. *Archiv der Mathematik*, 9(4): 389–393, 1958.

David Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272, 1953.

Francis Bloch and David Cantala. Dynamic assignment of objects to queuing agents. *American Economic Journal: Microeconomics*, 9(1):88–122, 2017.

Yeon-Koo Che and Olivier Tercieux. Optimal queue design. *Available at SSRN 3743663*, 2020.

Daniele Condorelli. What money can't buy: Efficient mechanism design with costly signals. *Games and Economic Behavior*, 75(2):613–624, 2012.

Laura Doval and Balász Szentes. On the efficiency of queueing in dynamic matching markets. Technical report, Working paper, Division of the Humanities and Social Sciences, California ..., 2019.

Piotr Dworczak, Scott Duke Kominers, and Mohammad Akbarpour. Redistribution through markets. *Econometrica*, 89(4):1665–1698, 2021.

Matthew Gentzkow and Emir Kamenica. A Rothschild-Stiglitz approach to bayesian persuasion. *American Economic Review*, 106(5):597–601, 2016.

Alex Gershkov, Benny Moldovanu, Philipp Strack, and Mengxi Zhang. A theory of auctions with endogenous valuations. *Journal of Political Economy*, 129(4):1011–1051, 2021.

Neil E Gretsky, Joseph M Ostroy, and William R Zame. Subdifferentiability and the duality gap. *Positivity*, 6(3):261–274, 2002.

G.H. Hardy, J.E. Littlewood, and G. Pólya. Some simple inequalities satisfied by convex functions. *Messenger of Mathematics*, 1929.

Jason D. Hartline and Tim Roughgarden. Optimal mechanism design and money burning. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pages 75–84. ACM, 2008.

Andreas Kleiner, Benny Moldovanu, and Philipp Strack. Extreme points and majorization: Economic applications. *Econometrica*, 89(4):1557–1593, 2021.

Anton Kolotilin. Optimal information disclosure: A linear programming approach. *Theoretical Economics*, 13(2):607–635, 2018.

Glen S Leithead. Model for bathtub-shaped hazard rate: Monte Carlo study. 1970.

Jacob D Leshno. Dynamic matching in overloaded waiting lists. *American Economic Review*, 112 (12):3876–3910, 2022.

David Lingenbrink and Krishnamurthy Iyer. Optimal signaling mechanisms in unobservable queues with strategic customers. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 347–347. ACM, 2017.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2): 583–601, 2002.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Afshin Nikzad. Multi-criteria mechanisms: Constraints and comparative statics. *Working paper*, 2023.

OPTN. Organ procurement and transplantation network policies, 2014.

OPTN. Ethical principles in the allocation of human organs, 2015. URL https://optn.transplant.hrsa.gov/professionals/by-topic/ethical-considerations/ethical-principles-in-the-allocation-of-human-organs/.

OPTN. Manipulation of the organ allocation system waitlist priority through the escalation of medical therapies, 2018. URL https://optn.transplant.hrsa.gov/media/2500/ethics_whitepaper_201806.pdf.

OPTN. A guide to calculating and interpreting the Kidney Donor Profile Index (KDPI). 2022.

Pietro Ortoleva, Evgenii Safonov, and Leeat Yariv. Who cares more? allocation with diverse preference intensities. Technical report, National Bureau of Economic Research, 2021.

Peter P Reese, Olivier Aubert, Maarten Naesens, Edmund Huang, Vishnu Potluri, Dirk Kuypers, Antoine Bouquegneau, Gillian Divard, Marc Raynaud, Yassine Bouatou, et al. Assessment of the utility of kidney histology as a basis for discarding organs in the united states: A comparison of international transplant practices and outcomes. *Journal of the American Society of Nephrology*, 32(2):397–409, 2021.

Eran Simhon, Yezekael Hayel, David Starobinski, and Quanyan Zhu. Optimal information disclosure policies in strategic queueing games. *Operations Research Letters*, 44(1):109–113, 2016.

Stanford R&DE. Applying for Stanford Graduate Housing Brochure, 2022. URL https://rde.stanford.edu/sites/default/files/Housing/PDF/2022-23_Grad_Brochure.pdf.

Xuanming Su and Stefanos A. Zenios. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Management Science*, 52(11):1647–1660, 2006. doi: 10.1287/mnsc.1060.0541. URL https://doi.org/10.1287/mnsc.1060.0541.

Tien Thanh Thach. A three-component additive weibull distribution and its reliability implications. *Symmetry*, 14(7):1455, 2022.

Neil Thakral. Matching with stochastic arrival. In *AEA Papers and Proceedings*, volume 109, pages 209–12, 2019.

Juuso Toikka. Ironing without control. *Journal of Economic Theory*, 146(6):2510–2526, 2011.

A Proof of Theorem 3.1

We first define some preliminaries and outline the proof steps. The full proof is presented afterward.

A.1 Preliminaries and proof steps

For a positive integer i, denote by $[i]$ the set $\{1,\dots,i\}$. The next proposition characterizes the waiting costs and queue lengths in an incentive-compatible monotone disjoint queue mechanism. Recall that the length of a queue equals the agents' arrival rate at that queue times the waiting time in that queue.

Proposition A.1. *Let $\overline{\omega}_i$ be the average object type sent to the queue q_i in an incentive-compatible monotone disjoint queue mechanism. Then, the steady-state waiting time at queue q_i is*

$$t_i = \begin{cases} c^{-1}(u(\theta_1,\overline{\omega}_1) - u(\theta_1,\omega_0)) & \text{for } i=1,\\ c^{-1}(u(\theta_i,\overline{\omega}_i) - u(\theta_i,\overline{\omega}_{i-1}) + c(t_{i-1})) & \text{for } i>1. \end{cases}$$

and the length of q_i is $\alpha_i t_i$, where α_i is the agents' arrival rate at q_i.

Proof. Let the monotone disjoint queue mechanism be defined by the sequence $\theta_0 = 0, \theta_1, \dots, \theta_k, \theta_{k+1} = \overline{\theta}$, as we defined in Section 3. The payoff of an agent of type θ from joining a queue q_i is then given by $\pi_i(\theta) = u(\theta,\overline{\omega}_i) - c(t_i)$. By the incentive compatibility constraint and the continuity of $\pi_1(\,)$, agents of type θ_1 are indifferent between joining queue q_1 and being assigned to the object ω_0. Thus, $u(\theta_1,\omega_0) = u(\theta_1,\omega_1) - c(t_1)$, which means $t_1 = c^{-1}(u(\theta_1,\omega_1) - u(\theta_1,\omega_0))$.

Similarly, by the continuity of $\pi_2(\,)$, agents of type θ_2 are indifferent between joining queues q_1 and q_2. Thus, $\pi_1(\theta_2) = \pi_2(\theta_2)$. That is, $u(\overline{\omega}_2,\theta_2) - c(t_2) = u(\overline{\omega}_1,\theta_2) - c(t_1)$, which gives

$$t_2 = c^{-1}(u(\overline{\omega}_2,\theta_2) - u(\overline{\omega}_1,\theta_2) + c(t_1)).$$

The same argument applies inductively to all $i \geq 2$, which proves the claim. □

Let $\mathsf{L}(\Theta)$ denote the set of all increasing Lebesgue integrable functions from Θ to $[\omega_0,\omega_n]$. We recall some definitions from *majorization* theory (Hardy et al., 1929; Kleiner et al., 2021). For nondecreasing $X, Y \in \mathsf{L}(\Theta)$, we say that X *majorizes* Y, denoted by $Y \prec X$, if

$$\int_\theta^{\bar\theta} Y(s)\,\mathrm{d}G(s) \leq \int_\theta^{\bar\theta} X(s)\,\mathrm{d}G(s) \text{ for all } \theta \in \Theta,$$
$$\int_0^{\bar\theta} Y(s)\,\mathrm{d}G(s) = \int_0^{\bar\theta} X(s)\,\mathrm{d}G(s).$$

We say that X *weakly majorizes* Y, denoted by $Y \prec_{\mathrm{w}} X$, if the first condition (but not necessarily

the second condition) holds. Define $\text{MPS}(X) = \{Y \in \Delta(\Theta) : Y \succeq X\}$. Define $\text{MPS}_w(X)$ in the same way but with $Y \succeq X$ replaced with $Y \succeq_w X$.

It is helpful to consider the interim allocation rule X_{PAM} corresponding to the positive assortative assignment. Formally, for $q \in [0, q)$, let $F^{-1}(q)$ denote the lowest object type $\omega \in \Omega$ such that the total arrival rate of all objects with type weakly lower than ω is (strictly) larger than q. Also, let $F^{-1}(1) = \omega_n$. Then, define

$$X_{\text{PAM}}(\theta) = \begin{cases} F^{-1}(G(\theta)) & \text{for } \theta \geq G^{-1}(1 - N), \\ \omega_0 & \text{for } \theta < G^{-1}(1 - N). \end{cases}$$

We will consider the reduced-form version of the planner's problem: choosing an interim allocation rule $X \in \mathcal{X}$ to maximize the planner's objective (3.1). We then will show that (i) the solution to the planner's problem is *nonwasteful*, meaning that it allocates all objects to agents, (ii) the interim allocation rules of all nonwasteful, incentive-compatible, and individually rational mechanisms belong to $\text{MPS}(X_{\text{PAM}})$, and (iii) the planner's objective is a convex functional in X. This fact, together with the following lemma, makes Bauer's Maximum principle applicable.

Lemma 0 (Proposition 1 in Kleiner et al. (2021))**.** *The sets $\text{MPS}(X_{\text{PAM}})$ and $\text{MPS}_w(X_{\text{PAM}})$ are convex and compact in the norm topology.*

From Bauer's Maximum Principle it follows that the planner's objective (3.1) attains its maximum at an extreme point of $\text{MPS}(X_{\text{PAM}})$. We then use the following extreme point characterization result of Kleiner et al. (2021), translated to our setting, to conclude that the interim allocation rule X that maximizes the planner's objective has an *interval structure*. In the final step of the proof we show that this structure corresponds to a monotone disjoint queue mechanism.

Theorem 0 (Theorem 1 of Kleiner et al. (2021))**.** *$X : \Theta \to [\omega_0, \omega_n]$ is a right-continuous extreme point of $\text{MPS}(X_{\text{PAM}})$ if and only if there is a family of disjoint intervals $[\underline{\theta}_i, \bar{\theta}_i)$ indexed by $i \in I$ such that*

$$X(\theta) = \begin{cases} X_{\text{PAM}}(\theta), & \theta \notin \bigcup_{i \in I} [\underline{\theta}_i, \bar{\theta}_i) \\ \frac{\int_{\underline{\theta}_i}^{\bar{\theta}_i} X_{\text{PAM}}(s)\, \mathrm{d}G(s)}{G(\bar{\theta}_i) - G(\underline{\theta}_i)}, & \theta \in [\underline{\theta}_i, \bar{\theta}_i). \end{cases}$$

Definition A.1. *We call the family of intervals in Theorem 0 a* family of intervals characterizing the extreme point X.

This completes the description of preliminaries and proof steps. The complete proof is below.

A.2 Proof of Theorem 3.1

Given a mechanism M, we use X_M to denote the interim allocation rule of the mechanism; thus $X_M(\theta)$ gives the expected object type assigned to an agent who reports θ to M. We use $c_M(\theta)$ to

denote the expected waiting cost incurred by an agent who reports type θ to M. We call (X_M, c_M) the *reduced form* of M. From standard arguments in mechanism design it follows that a necessary condition for a mechanism M to be incentive compatible is its interim allocation rule $X_M(\theta)$ being increasing and $c_M(\theta)$ being determined from X_M by the envelope condition (e.g., see Milgrom and Segal (2002)). For completeness, we include the statement below.

Lemma A.1. *Let u satisfy Assumption 2.1. Then, $(X_M(\theta), c_M(\theta))$ is the reduced form of an incentive-compatible and individually rational mechanism M only if $X_M(\theta)$ is increasing in θ and*

$$c_M(\theta) = u(\theta, X_M(\theta)) \quad \int_0^\theta u_1(\tau, X_M(\tau))\,\mathrm{d}\tau, \theta \; 2 \; \Theta. \tag{A.1}$$

Given a mechanism M as above, we next write the planner's objective as a functional of X_M. To this end, define the *virtual welfare* function $\phi : \Theta \quad [\omega_0, \omega_n] \; ! \; \mathsf{R}$ as

$$\phi(\theta, x) = \frac{1 \quad G(\theta)}{g(\theta)} \quad u_1(\theta, x).$$

Lemma A.2. *The planner's objective in an incentive-compatible and an individually rational mechanism M equals*

$$\int_0^{\bar\theta} \Big(u(\theta, X_M(\theta))\lambda_E + \phi(\theta, X_M(\theta))\lambda_W \Big)\,\mathrm{d}G(\theta).$$

Proof. First, we observe that efficiency under a mechanism M is $E_M = \int_0^{\bar\theta} u(\theta, X_M(\theta))\,\mathrm{d}G(\theta)$. Second, by (A.1) we can write the welfare under M as

$$\begin{aligned} W_M = \int_0^{\bar\theta} [u(\theta, X_M(\theta)) \quad c_M(\theta)]\,\mathrm{d}G(\theta) &= \int_0^{\bar\theta}\int_0^{\theta} u_1(\tau, X_M(\tau))d\tau\,\mathrm{d}G(\theta) \\ &= \int_0^{\bar\theta} \frac{1 \quad G(\theta)}{g(\theta)} u_1(\theta, X_M(\theta))\,\mathrm{d}G(\theta) = \int_0^{\bar\theta} \phi(\theta, X_M(\theta))\,\mathrm{d}G(\theta) \end{aligned}$$

where the second equality is by reversing the integration order due to Fubini's theorem. The proof is complete as the planner's objective is $E_M\lambda_E + W_M\lambda_W$. □

Lemma A.2 writes the planner's objective as a functional of the interim allocation rule X. Denote this functional by $\Pi(X)$. The support of Π is defined to be $\mathsf{L}(\Theta)$. For notational simplicity, we define $W(X) = \int_0^{\bar\theta} \frac{1 \; G(\theta)}{g(\theta)} u_1(\theta, X(\theta))\,\mathrm{d}G(\theta)$ to denote welfare and $E(X) = \int_0^{\bar\theta} u(\theta, X(\theta))\,\mathrm{d}G(\theta)$ to denote efficiency for a given interim allocation rule X.

Lemma A.3. *The functional $\Pi(X)$ is convex in X.*[35]

[35] That is, $\Pi(\alpha X + (1 \quad \alpha)Y) \quad \alpha\Pi(X) + (1 \quad \alpha)\Pi(Y)$ for all X, Y in the support of Π and $\alpha \; 2 \; (0, 1)$.

Proof. The welfare $W(X)$ is convex in X since u_1 is convex in its second argument. Since $u(\theta,\omega) = \int_0^\theta u_1(\kappa,\omega)\,\mathrm{d}\kappa$, and since $u_1(\kappa,\omega)$ is convex in its second argument for every κ 2 Θ, then so is $u(\theta,w)$. It follows that $E(X)$ is convex in X and, thus, so is $\Pi(X) = \lambda_W W(X) + \lambda_E E(X)$. □

Lemma A.4. *X* *MPS*$_{\mathrm{w}}(X_{\mathrm{PAM}})$.

Proof. We will show that if X 2 MPS$_{\mathrm{w}}(X_{\mathrm{PAM}})$, then X cannot be the interim allocation rule of an individually rational and incentive-compatible mechanism. The proof is by contradiction. Suppose that X does correspond to the interim allocation rule of an incentive-compatible and individually rational mechanism M. Let θ be an arbitrary agent type at which weak majorization fails; i.e.,

$$\int_\theta^{\bar\theta} X(s)\,\mathrm{d}G(s) > \int_\theta^{\bar\theta} X_{\mathrm{PAM}}(s)\,\mathrm{d}G(s).$$

Note that X_{PAM} allocates a mass $1 \quad G(\theta)$ of objects with an average quality $\overline{\omega} = \int_\theta^{\bar\theta} X_{\mathrm{PAM}}(s)\,\mathrm{d}G(s)$, whereas X allocates an equal mass of objects but with an average quality higher than $\overline{\omega}$. This is a contradiction, since X_{PAM} assigns objects to agents in a positive assortative manner. □

We next write a relaxed version of the planner's problem, such that the solution to the relaxed problem is an upper bound on the solution to the planner's problem. We then complete the proof by showing that the solution to the relaxed problem corresponds to a monotone disjoint queue mechanism. The relaxed problem that we consider is

$$\max_{X \,2\, \mathrm{MPS_w}(X_{\mathrm{PAM}})} \Pi(X). \tag{A.2}$$

This is a relaxation since *X* MPS$_{\mathrm{w}}(X_{\mathrm{PAM}})$, by Lemma A.4.

Lemma A.5. *Any optimal solution* X *to* (A.2) *satis es* X *2 MPS*(X_{PAM}).

Proof. The proof is by contradiction. Suppose X 62 MPS(X_{PAM}). Then, there is $\epsilon, \kappa > 0$ satisfying

$$\epsilon < \int_\theta^{\bar\theta} X_{\mathrm{PAM}}(s)\,\mathrm{d}G(s) \quad \int_\theta^{\bar\theta} X\ (s)\,\mathrm{d}G(s) \tag{A.3}$$

for all θ 2 $[0,\kappa]$ and $X\ (\kappa) < X_{\mathrm{PAM}}(\kappa)$. This holds since X 2 MPS$_{\mathrm{w}}(X_{\mathrm{PAM}})$ by Lemma A.4, and since both of the above integrals are continuous in θ. Define the function $Z : \Theta$! R_+ by

$$Z(\theta) = \begin{cases} X\ (\kappa) \quad X\ (\theta), & \text{if } \theta < \kappa \\ X\ (\theta), & \text{otherwise.} \end{cases}$$

Observe that $X \ + \delta Z$ is an increasing function for every δ 2 $[0,1]$. Choose $\delta > 0$ such that $X \ + \delta Z$ 2 MPS(X_{PAM}). This holds for any sufficiently small δ by (A.3). Since $X\ (\kappa) < X_{\mathrm{PAM}}(\kappa)$,

then $X + \delta Z$ is strictly larger than X at every point in $[0, \kappa)$. It follows that $\Pi(X + \delta Z) > \Pi(X)$, as u is strictly supermodular. This, however, contradicts the optimality of X. □

By Lemma A.5, any optimal solution X to (A.2) satisfies $X \in \mathrm{MPS}(X_{\mathrm{PAM}})$. By Lemma 0 $\mathrm{MPS}(X_{\mathrm{PAM}})$ is convex and compact in the norm topology. On the other hand, $\Pi(X)$ is a convex functional, by Lemma A.3. Thus, Bauer's Maximum principle is applicable to (A.2), and implies that there is an optimal solution to this problem that is an extreme point of $\mathrm{MPS}(X_{\mathrm{PAM}})$. We choose X to be such a solution and, without loss of generality, right-continuous. We next show that such X can be transformed to a monotone disjoint queue mechanism M such that $\lambda_E E_M + \lambda_W W_M = \Pi(X)$. This would mean that M is optimal.

By Theorem 0, X is characterized by a collection of disjoint intervals $[\underline{\theta}_i, \overline{\theta}_i)$ indexed by $i \in I$ such that X equals X_{PAM} outside of the intervals and inside each interval X equals the average of X_{PAM} over that interval. As X_{PAM} is piecewise constant with at most n jump discontinuities, then we can assume that $|I| \leq n$, without loss of generality. To see why, suppose there is $j \in I$ such that X_{PAM} is continuous over $[\underline{\theta}_j, \overline{\theta}_j)$. Then X_{PAM} must be constant over this interval. Thus, the average of X_{PAM} over this interval equals its value over the interval. This means that the extreme point characterized by the family of intervals $\mathrm{Inf}\{\}$ is the same as X. Hence, without loss of generality we can assume that X_{PAM} has at least one jump discontinuity at an interior point of every interval in I. Therefore, $|I| \leq n$. As the number of intervals in I is finite, we can also assume without loss of generality that they are ordered so that $\underline{\theta}_1 < \ldots < \underline{\theta}_{|I|}$. Let $\Theta_I = \{\theta : \exists i \in I \text{ such that } \theta \in [\underline{\theta}_i, \overline{\theta}_i)\}$.

We now construct the monotone disjoint queue mechanism M from X. The mechanism involves two types of queues: one distinct queue for each interval in I, and at most one distinct queue for every object type, as defined below:

i. For every $i \in I$, M contains a distinct queue $\hat{q}_i$. All agents of a type belonging to $[\underline{\theta}_i, \overline{\theta}_i)$ are sent to this queue. For every $\omega \in \Omega$, define $T_i(w) = \{\theta : \theta \in [\underline{\theta}_i, \overline{\theta}_i), X_{\mathrm{PAM}}(\theta) = \omega\}$. Let $R_i(\omega) = G(T_i(\omega))$, where with slight abuse of notation we denote by $G(S)$ the measure of a subset $S \subseteq \Theta$ of agent types. Objects of type ω are sent to $\hat{q}_i$ at rate $R_i(\omega)$, for all $\omega \in \Omega$.

ii. For every $w \in \Omega$, let $T^{\emptyset}(\omega) = \{\theta : \theta \in \Theta \cap \Theta_I, X_{\mathrm{PAM}}(\theta) = \omega\}$. Also, let $R^{\emptyset}(\omega) = G(T^{\emptyset}(\omega))$. If $R^{\emptyset}(\omega) > 0$, then M contains a distinct queue $\hat{\hat{q}}_\omega$ to which only objects of type ω are sent, and at rate $R^{\emptyset}(\omega)$. All agents with type belonging to $T^{\emptyset}(\omega)$ are sent to $\hat{\hat{q}}_\omega$.

We next verify that the allocation of objects and agents to queues in M are consistent with a monotone disjoint queue mechanism. The set of agent types sent to each queue is a convex subset (i.e., an interval) of agent types. For a queue q in M we say that q has type t if the expected object type sent to q equals t. Note that an agent of type θ is sent to a queue of type $X(\theta)$. Thus, two agents of types $\theta_1, \theta_2 \in \Theta$ with $\theta_1 < \theta_2$ are sent to the same queue if and only if $X(\theta_1) = X(\theta_2)$. Otherwise, $X(\theta_1) < X(\theta_2)$, and the agent of type θ_1 is sent to a lower-type queue than the agent of type θ_2. It follows that the agent of type θ_2 may receive a lower object type than the agent of type θ_1 only if both agents are sent to the same queue in M. Hence, M is a monotone disjoint

queue mechanism. Moreover, its interim allocation rule is X . The steady-state waiting times (i.e., waiting costs) in the queues of M are uniquely determined by Proposition A.1, so that M is individually rational and incentive compatible. Hence, M achieves an objective value of $\Pi(X)$ as its interim allocation rule is X , and thus is optimal by Lemma A.2.

B Proofs for Section 4

Consider a steady-state equilibrium F of the FCFS waitlist with deferrals under a disclosure policy μ. We say F is *positive assortative with respect to* μ if there are no two agents in F of distinct types such that the higher-type agent is assigned to an object with a lower interim type.

B.1 Proof of Proposition 4.1

For notational simplicity, let $\Omega_\mu = \{\alpha_1, \ldots, \alpha_m\}$ with $\alpha_1 > \ldots > \alpha_m$. If $F_\mu(\alpha_1) \quad 1$ then there is a unique steady-state equilibrium in which every agent is allocated immediately upon arrival an object with interim type α_1. Thus, suppose $F_\mu(\alpha_1) < 1$. Define

$$t_1 = \sup_{t \in T} \{t : \exists \theta \in \Theta \text{ such that } a(t, \theta) > 0\}.$$

First, we will show that all objects of type α_1 are allocated at time t_1.

Claim B.1. *$M(t_1, \alpha_1) = F_\mu(\alpha_1)$, and $M(t_1, \alpha_i) = 0$ for all $i > 1$.*

Proof. The proof for $M(t_1, \alpha_1) = F_\mu(\alpha_1)$ is by contradiction. Suppose $M(t_1, \alpha_1) < F_\mu(\alpha_1)$. Then, there is no $t' < t_1$ such that $\int_{\theta \in \Theta} a(t', \theta) \mathrm{d}\theta = 0$, because otherwise $a(t_1, \theta) = 0$ for all $\theta \in \Theta$, which is a contradiction. Also, there is $\epsilon > 0$ such that for all $t \in (t_1 \quad \epsilon, t_1)$, $b(t, \alpha_1) > 0$. The reason is that, otherwise, $b(t', \alpha_1) = 0$ for all $t' < t$, which implies that $M(t_1, \alpha_1) = F_\mu(\alpha_1)$ as b is increasing in its first argument.

It follows that for all $t \in (t_1 \quad \epsilon, t_1)$, $M(t, \alpha_1) > 0$. Thus, for every such t, there is a jump discontinuity in $b(\,, \alpha_1)$ at point t. This, however, contradicts the fact that $b(\,, \alpha_1)$ is monotone, since any monotone function has countably many jump discontinuities, by Froda's theorem.

Next we prove the claim for $i > 1$. The proof is by contradiction. Suppose $M(t_1, \alpha_i) > 0$ for some $i > 1$. Then, is $\theta \in \Theta$ such that $m(t_1, \theta, \alpha_i) > 0$. This contradicts the optimality of the decision rule D_θ for agents of type θ; such an agent can increase her payoff by switching to a new decision rule D'_θ defined by

$$D'_\theta(t) = \begin{cases} D_\theta(t) & \text{for } t \quad t_1 \quad \delta, \\ \alpha_1 & \text{for } t > t_1 \quad \delta, \end{cases}$$

where δ is a sufficiently small positive constant. □

Repeating the same argument as the proof of Claim B.1 shows that there is a nonnegative $t_2 < t_1$ such that $M(t_2, \alpha_2) = F_\mu(\alpha_2)$, and $M(t, \alpha_i) = 0$ for all $t > t_2$ and i 2. Inductively, this argument shows that for every integer j m satisfying $\sum_{l=1}^{j} F_\mu(\alpha_l) < 1$, there exists $t_j > 0$ such that $t_j < t_j$ 1, $M(t_j, \alpha_j) = F_\mu(\alpha_j)$, and $M(t, \alpha_i) = 0$ for all $t > t_j$ and i j. Let j be the smallest integer satisfying $\sum_{l=1}^{j} F_\mu(\alpha_l)$ 1. It follows that every agent that does not allocate an object with interim type larger than α_j allocates an object with interim type α_j upon her arrival.

Therefore, in any steady-state equilibrium of the FCFS policy, we have a *positive assortative assignment with respect to* μ, i.e., there are no two agents of types θ, θ' with $\theta > \theta'$ such that the interim object assigned to the agent of type θ is lower than the interim object type assigned to the agent of type θ'. (Otherwise the agent of type θ is better off if she chooses the decision rule of the other agent, due to the supermodularity of u.) This proves the second claim of the proposition.

To prove the first claim (uniqueness), consider an agent of type θ and a decision rule that sets $D_\theta(t) = \alpha_j$ for all t t_j. This decision rule guarantees that the agent allocates an object with interim type α_j at time t_j, and thus a payoff of $u(\theta, \alpha_j)$ $c(t_j)$. Also, since ω_0 2 Ω, then $\sum_{x 2 \Omega_\mu} F_\mu(x) = 1 + N$, which is larger than the arrival rate of agents 1. As agents with a type sufficiently close to 0 prefer not to wait for any object and to be allocated immediately, then the waiting time for obtaining the object with the lowest interim type α_m must equal 0; i.e., $t_m = 0$.

As any steady-state assignment is positive assortative with respect to μ, there are decreasingly ordered agent types $\theta_0 = \bar{\theta}, \theta_1, \ldots, \theta_m = 0$ such that agents of a type in $(\theta_i, \theta_i$ 1) are allocated α_i, for all i. Moreover, type θ_i should attain the same payoff from the decision rules $D_{\theta_i'}, D_{\theta_i''}$ where θ_i' 2 $(\theta_i, \theta_i$ 1) and θ_i'' 2 (θ_{i+1}, θ_i). We call this the *indi erence condition* for θ_i. The indifference condition for θ_m, together with $t_m = 0$, uniquely determines t_m 1. Inductively, the indifference condition for type θ_i together with t_i uniquely determines t_i 1, for $i > 1$. Thus, the interim object type allocated and the expected waiting time of an agent of type θ 2 $\Theta n f \theta_i g_{i=0}^{m}$ is the same across all steady-state equilibria. This means that the FCFS waitlist with deferrals under μ has an essentially unique steady-state equilibrium.

B.2 Proof of Theorem 4.1

Consider an incentive-compatible and individually rational mechanism M. We say a disclosure policy μ *implements* M if the following holds: There is a finite subset of agent types Θ' Θ so that, for any θ 2 $\Theta n \Theta'$ and in any steady-state equilibrium of the FCFS waitlist with deferrals under μ, (i) the expected object type allocated to any agent of type θ is is the same as the expected object type allocated to θ under M, and (ii) the expected waiting cost incurred by any agent of type θ is is the same as the expected waiting cost incurred by that agent type under M. We note that the expectations are computed at an agent's arrival time.

Lemma B.1 (Indirect Implementation)**.** *For any disjoint queue mechanism M, there is a disclosure policy μ_M that implements M.*

Proof. Let $q_0, q_1, \ldots, q_k$ denote the queues used in M. Also, let $\overline{\omega}_i$ be the expected type of an object conditional on the object being sent to q_i. Without loss of generality, we assume that $\overline{\omega}_i \neq \overline{\omega}_j$ for $i \neq j$. This is without loss of generality because two queues q_i, q_j with $\overline{\omega}_i = \overline{\omega}_j$ can be merged into one queue $q_{i,j}$, such that (i) the the set of objects sent to $q_{i,j}$ is the union of the set of objects sent to q_i and q_j, and (ii) the set of agents sent to $q_{i,j}$ is the union of the set of agents sent to q_i and q_j. By incentive compatibility, the waiting times at q_i, q_j are equal. The waiting time at $q_{i,j}$ would be the same. The payoff of every agent who joins $q_{i,j}$ thus equals her payoff from joining q_i or q_j. Thus, without loss of generality we suppose that the queues are labeled such that $\overline{\omega}_0 < \overline{\omega}_1 < \ldots < \overline{\omega}_k$.

Let $r_{\omega,i}$ be the steady-state rate of objects of type $\omega \in \Omega$ that are assigned to q_i in M. We next construct the disclosure policy μ_M, which consists of a finite realization space $\widehat{\Omega}$ and a family of distributions $\{\mu_M(\cdot|\omega)\}_{\omega \in \Omega}$ over $\widehat{\Omega}$. For each queue q_i with $i \geq 0$, $\widehat{\Omega}$ contains a signal realization s_i, i.e., $\widehat{\Omega} = \{s_0, \ldots, s_k\}$. For $i \geq 1$, define $\mu_M(\cdot|\omega)$ to be the probability distribution that assigns probability mass $\frac{r_{\omega,i}}{\sum_{j=0}^{k} r_{\omega,j}}$ to the signal s_i, for $s_i \in \widehat{\Omega}$ and $\omega \in \Omega$. Also, let $\mu_M(\cdot|\omega_0)$ be the degenerate distribution with support s_0.

Observe that $\Omega_{\mu_M} = \{\overline{\omega}_0, \ldots, \overline{\omega}_k\}$. By Proposition 4.1, the FCFS waitlist with deferrals under μ_M has an essentially unique steady-state equilibrium. Let E denote a steady-state equilibrium. Also, recall that E is positive assortative with respect to μ_M. With slight abuse of notation, we let $X_E(\theta), c_E(\theta)$ respectively denote the expected object type allocated to and the waiting cost incurred by an agent of type θ in E.

Let (X_M, c_M) be the reduced-form representation of M. Since M is an incentive-compatible disjoint queue mechanism, there is $\theta_0, \ldots, \theta_k \in \Theta$, with $0 = \theta_0 < \theta_1 < \ldots < \theta_k = \overline{\theta}$, such that $X_M(\theta) = \overline{\omega}_i$ holds for all $\theta \in (\theta_i, \theta_{i+1})$ and nonnegative integers $i < k$. As E is positive assortative with respect to μ, then $X_E(\theta) = X_M(\theta)$ for all $\theta \in (\theta_i, \theta_{i+1})$. We next show that $c_E(\theta) = c_M(\theta)$ for such θ as well. Briefly put, this is by the envelope theorem, since (X_E, c_E) and (X_M, c_M) satisfy the same set of incentive compatibility constraints. We include a direct proof below for completeness.

Let θ_i^0 be an arbitrary element of (θ_i, θ_{i+1}) for all $i < k$. The *indifference* condition for an agent of type θ_i is that $u(\theta_i, X_E(\theta_{i-1}^0)) - c_E(\theta_{i-1}^0) = u(\theta_i, X_E(\theta_i^0)) - c_E(\theta_i^0)$ holds when $c_E(\theta_i^0) > 0$. We will use this condition to show that $c_E(\theta_i^0) = c_M(\theta_i^0)$ for all i. The proof is by induction. The induction base case is all $i = 0$. By the individual rationality of M, $c_M(\theta_i^0) = 0$. Also, as shown in the proof of Proposition 4.1, $c_E(\theta_i^0) = 0$. Thus, $c_M(\theta) = c_E(\theta_i^0) = 0$ in this case. For the induction step, consider $i > 0$ with $c_E(\theta_{i+1}^0) > 0$, and suppose $c_E(\theta_j^0) = c_M(\theta_j^0)$ for all nonnegative integers $j \neq i$. We will show that the latter equality also holds for $j = i+1$. By the indifference condition for θ_{i+1},

$$\begin{aligned} c_E(\theta_{i+1}^0) &= u(\theta_{i+1}, X_E(\theta_{i+1}^0)) - u(\theta_{i+1}, X_E(\theta_i^0)) + c_E(\theta_i^0) \\ &= u(\theta_{i+1}, X_M(\theta_{i+1}^0)) - u(\theta_{i+1}, X_M(\theta_i^0)) + c_M(\theta_i^0) = c_M(\theta_{i+1}^0), \end{aligned}$$

where the second equality follows from the induction assumption and the third one from the indifference condition for θ_{i+1} in M. This completes the induction step and the proof of the theorem. □

Choose M be an optimal monotone disjoint queue mechanism (such as the one identified in Theorem 3.1). Suppose M involves k queues. By Lemma B.1, there is a disclosure policy μ_M that implements M. Recall from the proof that $\mu_M(\cdot|\omega)$ assigns probability mass $\frac{r_{\omega,i}}{\sum_{j=0}^{k} r_{\omega,j}}$ to the signal s_i, for $s_i \in \widehat{\Omega}$ and $\omega \in \Omega$. From the fact that M is a monotone disjoint queue mechanism, it follows that there are no two objects such that μ_M sends a lower signal realization for the object with higher type. That is, for $\omega_a, \omega_b \in \Omega$ with $\omega_a < \omega_b$, and $s_i, s_j \in \widehat{\Omega}$ with $i < j$, if $\mu_M(s_j|\omega_a) > 0$ then $\mu_M(s_i|\omega_b) = 0$. This property implies that the realization space of μ_M can be relabeled so that it pools adjacent object types, as follows. For a signal realization s_i, define $s_i^+ = \max\{\omega : \mu(s_i|\omega) > 0\}$ and $s_i^- = \min\{\omega : \mu(s_i|\omega) > 0\}$. Let μ be the same disclosure policy as μ_M, with the difference that every signal realization s_i is relabeled as $[s_i^-, s_i^+]$ in μ. Observe that μ satisfies the conditions of Definition 4.1, and thus pools adjacent object types.

C Proof of Theorem 5.1

The proof translates one of the findings of Nikzad (2023) from a mechanism design setting to an information design setting. We include the full proof here for completeness. The main step in the analysis is proving Theorem C.1. This theorem is used later in Section C.3 to prove Theorem 5.1.

Fact C.1. *There exist $\underline{e}, \bar{e} \in \mathbb{R}_+$ and a continuous decreasing concave function $w : [\underline{e}, \bar{e}] \to \mathbb{R}_+$ such that the efficiency-welfare Pareto frontier P equals the set of points $(e, w(e))$ for all $e \in [\underline{e}, \bar{e}]$.*

Proof. Recall that the set of interim allocation rules of all incentive-compatible and individually rational mechanisms, $\mathrm{MPS}_{\mathrm{w}}(X_{\mathrm{PAM}})$, is compact and convex by Lemma 0. Since efficiency and welfare both are linear functionals in the interim allocation rule, then M is convex and compact. If P is singleton, then the claim is proved. Otherwise, consider $(e_1, w_1), (e_2, w_2) \in P$ with $e_1 < e_2$. Then, it must hold that $w_1 > w_2$. For every $\alpha \in (0,1)$ and $e_3 = \alpha e_1 + (1-\alpha)e_2$, there is a point $(e_3, w_3) \in P$ by the convexity of M. Moreover, the unique line segment connecting the points (e_1, w_1) and (e_2, w_2) is contained in M. Thus, $w_3 \geq \alpha w_1 + (1-\alpha)w_2$, which proves the claim. □

We say that the Pareto frontier is *affine* at $e \in \mathrm{supp}(w)$ if w is an affine function over an open interval containing e. The Pareto frontier is *left-affine* at e if w is affine over an interval $(e', e]$. The Pareto frontier is *right-affine* at e if w is affine over an interval $[e, e')$.

A disclosure policy μ *generates* a point $(e, w) \in \mathbb{R}_+^2$ if, in the steady-state equilibrium of a FCFS waitlist with deferrals under μ, the average of the agents' utilities equals e and the average of the agents' payoffs equals w. (Recall that these averages depend only on μ and are the same across all steady-state equilibria, by Proposition 4.1.)

Definition C.1 (Relabeling of a realization space)**.** *Consider a disclosure policy μ with realization space $\widehat{\Omega}$. A disclosure policy μ'* equals μ up to a relabeling of the realization space *if the realization*

space of μ', denoted by $\widehat{\Omega}'$, has the same size as $\widehat{\Omega}$ and there is a one-to-one mapping $f : \widehat{\Omega} \to \widehat{\Omega}'$ such that $\mu'(f(s)|\omega) = \mu(s|\omega)$ for all $\omega \in \Omega$ and $s \in \widehat{\Omega}$.

Given a disclosure policy μ, if there is a disclosure policy μ' that pools adjacent object types and equals μ up to a relabeling of the realization space, then we say μ *pools adjacent object types up to a relabeling of its realization space*.

Theorem C.1. *The following holds:*

i. For an extreme point (e, w) of the Pareto frontier, there is a unique nondegenerate disclosure policy $\mu(e)$ that generates (e, w), where the uniqueness is up to a relabeling of the realization space; moreover, $\mu(e)$ pools adjacent object types.

ii. For a non-extreme point (e', w') of the Pareto frontier, let (e_l, e_r) be the maximal open interval containing e' such the Pareto frontier is a line over that interval. Then, every disclosure policy that generates (e', w') is more informative than $\mu(e_l)$ and less informative than $\mu(e_r)$. Define $\mu(e')$ to be the disclosure policy that sends a signal realization using $\mu(e_l)$ with probability $\frac{e_r - e'}{e_r - e_l}$, and otherwise sends a signal realization using $\mu(e_r)$. Then, $\mu(e')$ generates (e', w').

iii. For all $e, e' \in [\underline{e}, \overline{e}]$ with $e < e'$, $\mu(e)$ is less informative than $\mu(e')$.

We next use Figure 5 to briefly discuss the theorem, and then present its proof. The theorem implies that the point with the highest welfare on the Pareto frontier (point 1 in Figure 5) is generated by the least informative disclosure policy, and the point with the highest efficiency (point 6) is generated by the most informative one. Moreover, the points in between are monotonically ordered in terms of the informativeness of the disclosure policies that generate them. More precisely, recall that an extreme point of the Pareto frontier is a point that cannot be written as a convex combination of two other points on the Pareto frontier. For example, all of the marked points but point 4 in Figure 5 are extreme points. Thus, for $i \in \{1, 2, 3, 5, 6\}$, there is a unique nondegenerate disclosure policy η_i that generates point i. This holds by part (i) of the theorem. By part (ii) of the theorem, every disclosure policy that generates point 4 is more informative than η_3, and less informative than η_5. In particular, point 4 can be generated by a disclosure policy η_4 that randomizes over η_3 and η_5. This, together with part (i), implies that η_i is less informative than η_{i+1} for all $i < 6$. A similar argument shows that all of the points on the Pareto frontier are ordered in terms of informativeness, in the sense of part (iii) of the theorem (Figure 6).

We prove the theorem in Section C.2, after stating some preliminary definitions and results.

C.1 Proof of Theorem C.1: Preliminaries

We first state an inequality by Fan and Lorentz.

Theorem C.2 (Fan and Lorentz; 1954)**.** *Let $K : [0,1] \times [0,1] \to \mathbb{R}$. Then, $\int_0^1 K(x(q), q)dq \geq \int_0^1 K(y(q), q)$ holds for any two nondecreasing functions $x, y : [0,1] \to [0,1]$ such that $x \succeq y$ if and only if $K(p, q)$ is convex in p and supermodular in p, q.*

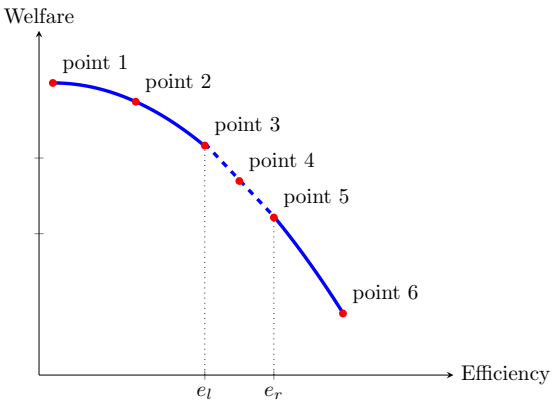


Figure 5: Point 4 is an interior point of an affine part of the Pareto frontier (the dotted segment), and thus not an extreme point; every other marked point is an extreme point.

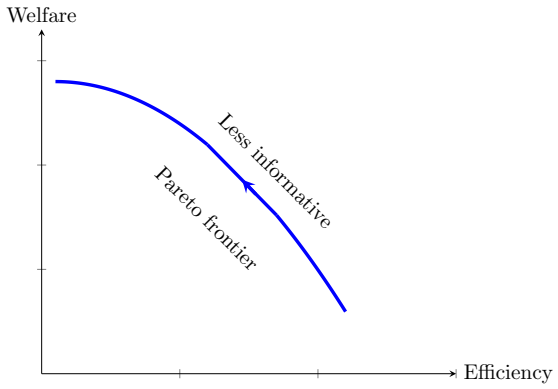


Figure 6: The disclosure policies implementing the Pareto frontier are ordered in terms of informativeness.

Interim allocation rules. We say X *2* MPS(X_{PAM}) is *canonical* if X is right-continuous and its range has a finite size.

Given a disclosure policy μ, we denote the interim allocation rule *associated* with μ by X_μ, and define it as follows. By Proposition 4.1, the FCFS waitlist with deferrals under μ has an essentially unique steady-state equilibrium. Consider a steady-state equilibrium F of the FCFS waitlist with deferrals under μ. Let $X_F(\theta)$ be the interim object type allocated to an agent of type θ in F. We choose F so that X_F is right-continuous. This is possible because we can choose F to be the steady-state equilibrium where agents who are indifferent between two acceptance thresholds choose the higher one. (There is in fact a unique steady-state equilibrium F that satisfies this property, by the proof of Proposition 4.1.) We define $X_\mu = X_F$.

Disclosure policies. For X *2* MPS(X_{PAM}), we say a disclosure policy μ *implements* X if there is a steady-state equilibrium of FCFS with deferrals under μ such that every agent of type θ in that equilibrium is allocated an object with interim type $X(\theta)$.

The disclosure policy $\mu[X]$. Consider a canonical function X *2* MPS(X_{PAM}). Let $\theta_0 = 0, \theta_1, \ldots, \theta_{k+1} = \bar{\theta}$ be such that X is constant over the interval $[\theta_i, \theta_{i+1})$ for all nonnegative integers i k. By $D[X]$ we denote the distribution that assigns a probability mass of $G(\theta_{i+1})$ $G(\theta_i)$ to $X(\theta_i)$, for all nonnegative integers i k. A distribution D that has finite support is the distribution of posterior means induced by some information disclosure policy if and only if $D[X_{\text{PAM}}]$ is a mean-preserving spread of D (Blackwell, 1953; Gentzkow and Kamenica, 2016; Kolotilin, 2018). Since X is majorized by X_{PAM}, then $D[X_{\text{PAM}}]$ is a mean-preserving spread of $D[X]$. This follows from the definition of $D[\,]$. The two latter facts imply that there is a nondegenerate disclosure policy that induces a distribution of posterior means $D[X]$. Let $\mu[X]$ denote such a disclosure

policy. If there are multiple such policies, let $\mu[X]$ denote an arbitrary one.

Lemma C.1. *If X 2 MPS(X_{PAM}) is canonical, then (i) the disclosure policy $\mu[X]$ implements X and generates $(E(X), W(X))$. Moreover, if X is an extreme point of MPS(X_{PAM}), then (ii) every nondegenerate disclosure policy that implements X equals $\mu[X]$ up to a relabeling of the realization space, and (iii) the disclosure policy $\mu[X]$ pools adjacent object types.*

Proof. Let $\theta_0 = 0, \theta_1, \ldots, \theta_{k+1} = \bar{\theta}$ be such that X is constant over the interval $[\theta_i, \theta_{i+1})$ for all nonnegative integers i k. Define $S = fX(\theta_i) : 0$ i kg. By the definition of $\mu[X]$, its realization space is the support of $D[X]$, which equals the set S.

To simplify notation, let $S = fs_0, \ldots, s_kg$ and denote $\mu[X]$ by μ. Consider a steady-state equilibrium F of the FCFS waitlist with deferrals under μ. Let $X_F(\theta)$ be the interim object type allocated to an agent of type θ in F, and let $c_F(\theta)$ be her expected waiting cost. We choose F so that X_F is right-continuous. This is possible because we can choose F to be the steady-state equilibium where agents who are indifferent between two acceptance thresholds choose the higher one. Since the realization space of μ is S, then $X_F(\theta)$ 2 S for all agent types θ. Also, F is positive assortative with respect to μ, by Proposition 4.1. It follows that $X_F(\theta) = X(\theta)$, for every θ 2 (θ_i, θ_{i+1}) and nonnegative integer i k. Hence, efficiency in F equals $E(X)$. It remains to show that welfare in F equals $W(X)$.

Observe that X_F, c_F satisfy the set of incentive-compatibility constraints $\theta X_F(\theta)$ $c_F(\theta)$ $\theta X_F(\theta')$ $c_F(\theta')$ for all θ, θ' 2 Θ. (Otherwise, an agent of type θ can benefit from adopting the strategy of an agent of type θ' in F, which contradicts with F being a steady-state equilibrium). Then, the envelope condition (A.1) implies that the waiting cost for an agent of type θ is $\theta X_F(\theta)$ $\int_0^\theta X_F(\tau)\,\mathrm{d}\tau$ in F. It follows that welfare equals $W(X)$ in F. By Proposition 4.1, the FCFS waitlist with deferrals under μ has an essentially unique steady-state equilibrium. Thus, welfare in every steady-state equilibrium equals $W(X)$. This proves part (i).

To prove part (ii), let η be an arbitrary nondegenerate disclosure policy that implements X. Thus, there is a steady-state equilibrium F' of the FCFS waitlist with deferrals under η such that the interim object type allocated to every agent of type θ in F' equals $X(\theta)$. Then, since η is nondegenerate, its realization space of η contains $k+1$ elements, $fs_0', \ldots, s_k'g$. Since X is an extreme point, then there is an object type ω_k such that $\mu(s_k j\omega) = \eta(s_k' j\omega) = 1$ for every $\omega > \omega_k$, and $\mu(s_k j\omega) = \eta(s_k j\omega) = 0$ for every $\omega < \omega_k$. Therefore, $\mu(s_k j\omega_k) = \eta(s_k' j\omega_k)$ holds because $\sum_{w2\Omega} \mu(s_k j\omega) = \sum_{w2\Omega} \eta(s_k' j\omega) = G(\theta_k)$ $G(\theta_k \ 1)$. Thus, $\mu(s_k j\omega) = \eta(s_k' j\omega)$ for all ω 2 Ω. Inductively, the same argument shows that $\mu(s_j j\omega) = \eta(s_j' j\omega)$ for all ω 2 Ω and every nonnegative integer $j < k$. Hence, η equals μ up to a relabeling of the realization space.

We next show that μ pools adjacent object types, up to a relabeling of its realization space. Let

$$\Omega_i = fX_{\text{PAM}}(\theta) : \theta \text{ 2 } [\theta_i, \theta_{i+1})g, \delta_{2i} = \min f\omega : \omega \text{ 2 } \Omega_i g, \delta_{2i+1} = \max f\omega : \omega \text{ 2 } \Omega_i g, \text{ and } s_i'' = [\delta_{2i}, \delta_{2i+1}].$$

We relabel the realization space of μ by replacing s_i with s_i^{00} in the realization space. We show that the resulting policy pools adjacent object types by verifying that it satisfies the conditions of Definition 4.1. First, a signal realization $[\delta_{2i}, \delta_{2i+1}]$ is realized only if the object type belongs to $[\delta_{2i}, \delta_{2i+1}]$. Second, every two intervals corresponding to two distinct signal realizations have at most one point in common, as X_{PAM} is increasing. Third, for ω 2 Ω, there is some θ with $X_{\text{PAM}}(\theta) = \omega$, which means there is some j with ω 2 $[\delta_{2j}, \delta_{2j+1}]$. This concludes the proof. □

Lemma C.2. *Let X_1, X_2 2 MPS(X_{PAM}) be canonical. If X_1 majorizes X_2, then any disclosure policy that implements X_1 is more informative than any disclosure policy that implements X_2. In particular, $\mu[X_1]$ is more informative than $\mu[X_2]$.*

Proof. Since X_{PAM} is piecewise constant, Theorem 0 implies that X_i is piecewise constant as well, for $i = 1, 2$. Thus, X_i is canonical. Consider a disclosure policy μ_i that implements X_i, for $i = 1, 2$. Then, $D[X_i]$ gives the distribution of posterior means induced by μ_i. Since X_1 majorizes X_2, then $D[X_1]$ is a mean-preserving spread of $D[X_2]$. It follows that μ_1 is more informative than μ_2. The last claim in the lemma holds, since $\mu[X_i]$ implements X_i by Lemma C.1. □

Remark C.1 (Informativeness of disclosure policies that pool adjacent object types)**.** Suppose that disclosure policies μ and η pool adjacent object types and respectively implement X, Y 2 MPS(X_{PAM}). From the interval structure defining μ, η it follows that X, Y are extreme points of MPS(X_{PAM}), by Theorem 0. If μ is more informative than η, then the distribution of posterior means induced by μ is a mean-preserving spread of that induced by η. Thus, X majorizes Y. Hence, the family of intervals characterizing Y is *coarser* than the family of intervals characterizing X; i.e., for every interval i in X, there is an interval in Y that contains i.

C.2 Proof of Theorem C.1

Let L_1 be the set of all bounded and integrable functions defined on the right-open unit interval. Functions y and z *coincide* over an interval if $z(q) = y(q)$ for every q belonging to that interval. We say a function z 2 L_1 is *subdi erentiable* at x_0 if there is a slope s 2 R such that $z(x)$ $z(x_0) + (x \quad x_0)s$ for all x. The slope s is also called a *subgradient* of z at x_0. The *supporting line for z at x_0* is the line $f(x, y) : y = z(x_0) + (x \quad x_0)sg$. We also say that this line *supports z at x_0*.

Let $\overline{z}$ denote the convex envelope of a real function z, i.e., the largest convex function that lies below z. An interval (α, β) is a *primitive ironed interval* of z if it is a maximal open interval with $z(q) > \overline{z}(q)$ for all q 2 (α, β). An interval (α, β) is a *maximal ironed interval* of z if it is a maximal open interval such that $\overline{z}(q)$ is affine over (α, β). (Figure 7 demonstrates these definitions.)

Lemma C.3. *Let a, z 2 L_1 be di erentiable and increasing and let a be convex. De ne $z_\epsilon(q) = z(q) + \epsilon a(q)$ for all q in the support of z and ϵ 2 R_+. (i) If z is subdi erentiable at x 2 $(0, 1)$, then so is z_ϵ. When a is strictly increasing, then the following holds: (ii) if z is subdi erentiable at q,*

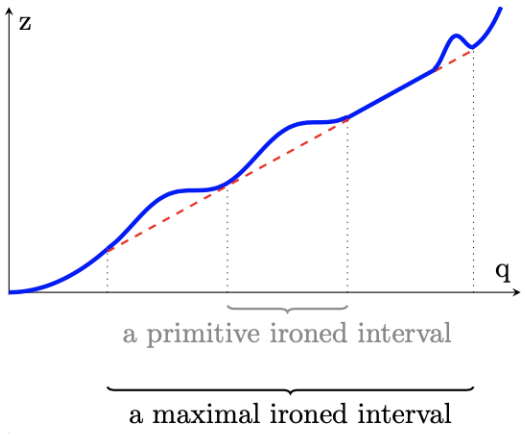


Figure 7

then, for $p < q$, there is no line that supports z_ϵ at both p and q, and (iii) every maximal ironed interval of z_ϵ is contained in a primitive ironed interval of z.

Proof. Since z is differentiable and subdifferentiable at x, then $z^0(x)$ is the subgradient of z at x. To prove claim (i), we show that $z^0_\epsilon(x)$ is a subgradient of z_ϵ at x. The reason is that

$$\begin{aligned} z_\epsilon(q) &= z(q) + \epsilon a(q) \quad z(x) + (q \quad x)z^0(x) + \epsilon a(q) \\ &= z(x) + (q \quad x)(z^0(x) + \epsilon a^0(x)) + \epsilon(a(q) \quad a^0(x)(q \quad x)) \\ &= z_\epsilon(x) + (q \quad x)z^0_\epsilon(x) + \epsilon(a(q) \quad a(x) \quad a^0(x)(q \quad x)) \quad z_\epsilon(x) + (q \quad x)z^0_\epsilon(x). \end{aligned}$$

The proof for claim (ii) is by contradiction. Suppose there is such a line l. Then,

$$\overline{z_\epsilon}(p) = z_\epsilon(p) = z(p) + \epsilon a(p) \quad z(q) \quad (q \quad p)z^0(q) + \epsilon a(p),$$

where the equalities hold because l supports z_ϵ at p and the inequality is the first-order lower bound for $z(p)$. On the other hand, since l supports z_ϵ at both p, q, then

$$\overline{z_\epsilon}(p) = z_\epsilon(q) \quad z^0_\epsilon(q)(q \quad p) = z(q) + \epsilon a(q) \quad (z^0(q) + \epsilon a^0(q))(q \quad p).$$

Combining the above two displayed equations, we get the following contradiction:

$$\begin{aligned} & z(q) + \epsilon a(q) \quad (z^0(q) + \epsilon a^0(q))(q \quad p) \quad z(q) \quad (q \quad p)z^0(q) + \epsilon a(p) \\ &) \; a(q) \quad (q \quad p)a^0(q) \quad a(p) \;) \; p = q. \end{aligned}$$

□

It remains to prove claim (iii). For the sake of contradiction, suppose the claim does not hold. Then, there is a maximal ironed interval (α, β) of z_ϵ and q *2* (α, β) such that z is subdifferentiable

at q. Then, z_ϵ is subdifferentiable at q, by claim (i) of the lemma. Let l be the line that supports z_ϵ at q. Since (α, β) is a maximal ironed interval of z_ϵ, then l also supports z_ϵ at α. This, however, cannot happen, by claim (ii) of the lemma. Contradiction.

Lemma C.4. *Let $k : \Theta \; ! \; \mathsf{R}$ be a bounded integrable function. Then, there are (not necessarily) distinct right-continuous extreme point solutions $X \; , X \;$ to the problem*

$$\max_{X 2 MPS(X_{\text{PAM}})} \int_\theta k(\theta) X(\theta) dG(\theta) \tag{C.1}$$

such that, for any solution Y to this problem, $X \;$ majorizes Y, and Y majorizes X . In particular, $E(X \;) < E(Y)$ if $Y \; 6= X \;$ and $E(Y) < E(X \;)$ if $Y \; 6= X \;$. Also, $E(X \;) < E(X \;)$ if $X \; 6= X \;$.

Proof. To prove this lemma, it is helpful to transform the parameters from the type space to the quantile space. We do this by a change of variables $q = G(\theta)$, as follows. Define the functions $\mathsf{k}, \mathsf{X}_{\text{PAM}}$ by $\mathsf{k}(q) = k(G^{-1}(q)), \mathsf{X}_{\text{PAM}}(q) = X_{\text{PAM}}(G^{-1}(q))$, for all $q \; 2 \; [0,1]$. We also define the majorization relation $\;_{\text{Q}}$ in the quantile space in the following way. For $\mathsf{X}, \mathsf{Y} \; 2 \; \mathsf{L}_1$, we write $\mathsf{Y} \;_{\text{Q}} \mathsf{X}$ when

$$\int_q^1 \mathsf{Y}(s)\,\mathrm{d}s \quad \int_q^1 \mathsf{X}(s)\,\mathrm{d}s \text{ for all } q \; 2 \; [0,1], \text{ and } \int_0^1 \mathsf{Y}(s)\,\mathrm{d}s = \int_0^1 \mathsf{X}(s)\,\mathrm{d}s.$$

Also, define $\text{MPS}_{\text{Q}}(\mathsf{X}_{\text{PAM}}) = f\mathsf{X} \; 2 \; \mathsf{L}_1 : \mathsf{X} \;_{\text{Q}} \mathsf{X}_{\text{PAM}} g$. We then can rewrite (C.1) as

$$\max_{\mathsf{X} 2 \text{MPS}_{\text{Q}}(\mathsf{X}_{\text{PAM}})} \int_0^1 \mathsf{k}(q)\mathsf{X}(q)\mathrm{d}q. \tag{C.2}$$

Define $\mathsf{K}(q) = \int_0^q \mathsf{k}(\kappa)\mathrm{d}\kappa$, and let $\overline{\mathsf{K}}$ denote the convex envelope of K. Let $\mathsf{I} \;$ be the set of all maximal ironed intervals of K. Also, let $\mathsf{I} \;$ be the set of all primitive ironed intervals of K. Let X and $\mathsf{X} \;$ be the extreme points of $\text{MPS}_{\text{Q}}(\mathsf{X}_{\text{PAM}})$ characterized by the families of intervals $\mathsf{I} \;$ and $\mathsf{I} \;$, respectively. (Recall that these extreme points are given by Theorem 0.) For a solution Y to (C.2), we say *majorization binds* in Y at point $q \; 2 \; [0,1)$ if $\int_q^1 \mathsf{Y}(p)\mathrm{d}p = \int_q^1 \mathsf{X}_{\text{PAM}}(p)\mathrm{d}p$.

Claim C.1. *Any right-continuous solution Y to (C.2) must satisfy the following: majorization binds in Y at q for any q not belonging to a maximal ironed interval of K. Also, majorization binds in Y at the leftmost point of any maximal ironed interval of K.*

Proof. For all $\mathsf{Z} \; 2 \; \text{MPS}_{\text{Q}}(\mathsf{X}_{\text{PAM}})$ it holds that

$$\int_0^1 \mathsf{k}(q)\mathsf{Z}(q)\mathrm{d}q = \mathsf{K}(1)\mathsf{Z}(1) \quad \int_0^1 \mathsf{K}(q)\mathrm{d}\mathsf{Z}(q)$$
$$\overline{\mathsf{K}}(1)\mathsf{Z}(1) \quad \int_0^1 \overline{\mathsf{K}}(q)\mathrm{d}\mathsf{Z}(q) = \int_0^1 \overline{\mathsf{K}}^{\emptyset}(q)\mathsf{Z}(q)\mathrm{d}q \quad \int_0^1 \overline{\mathsf{K}}^{\emptyset}(q)\mathsf{X}_{\text{PAM}}(q)\mathrm{d}q, \tag{C.3}$$

where the equalities are by integration by parts. The first inequality holds since $\overline{\mathsf{K}}(q) \quad \mathsf{K}(q)$, and the second one follows from the Fan-Lorentz theorem (Theorem C.2), because $\mathsf{Z} \quad {}_{\mathrm{Q}}\mathsf{X}_{\mathrm{PAM}}$ and $\overline{\mathsf{K}}^{\emptyset}$ is nondecreasing, as $\overline{\mathsf{K}}$ is convex. From classical convexification arguments it follows that the two inequalities hold with equality for any solution $\mathsf{Z} = \mathsf{Y}$ to (C.2). E.g., this is shown by inequalities (15) and (16) of Kleiner et al. (2021).

The proof of the claim is by contradiction. Consider a solution Y and $p \in (0,1)$ not belonging to a maximal ironed interval of K such that majorization does not bind in Y at p. As p does not belong to a maximal ironed interval of K and $\overline{\mathsf{K}}$ is convex, then $\overline{\mathsf{K}}^{\emptyset}(p_l) < \overline{\mathsf{K}}^{\emptyset}(p_r)$ for every p_l, p_r with $p_l < p < p_r$. Since $\mathsf{Y} \quad {}_{\mathrm{Q}}\mathsf{X}_{\mathrm{PAM}}$ and majorization does not bind in Y at p, it follows from the Fan-Lorentz theorem that

$$\int_0^1 \overline{\mathsf{K}}^{\emptyset}(q)\mathsf{X}_{\mathrm{PAM}}(q)\mathrm{d}q > \int_0^1 \overline{\mathsf{K}}^{\emptyset}(q)\mathsf{Y}(q)\mathrm{d}q.$$

This, however, contradicts the optimality of Y: the left- and right-hand sides must be equal since Y is a solution to (C.2), as assumed earlier. The same argument proves the claim when p is the leftmost point of a maximal ironed interval of K. □

We now show that $\mathsf{X} \quad {}_{\mathrm{Q}}\mathsf{Y}$. Suppose not. By Claim C.1, there is an interval $[\alpha, \beta) \in \mathsf{I}$ and $\kappa \in (\alpha, \beta)$ such that $\int_\kappa^1 \mathsf{Y}(p)\mathrm{d}p < \int_\kappa^1 \mathsf{X}\ (p)\mathrm{d}p$. Also, majorization binds in Y at α and β. Since Y is increasing and X is constant over $[\alpha, \beta)$, then $\int_\alpha^1 \mathsf{Y}(p)\mathrm{d}p < \int_\alpha^1 \mathsf{X}\ (p)\mathrm{d}p$, which is a contradiction. Thus, $\mathsf{X} \quad {}_{\mathrm{Q}}\mathsf{Y}$.

We next show that $\mathsf{Y} \quad {}_{\mathrm{Q}}\mathsf{X}$. For the sake of contradiction, suppose there is a solution Y to (C.2) for which the claim does not hold. Thus, there is $\kappa \in (0,1)$ such that $\int_\kappa^1 \mathsf{X}\ (p)\mathrm{d}p < \int_\kappa^1 \mathsf{Y}(p)\mathrm{d}p$. Since X is the extreme point corresponding to I , then κ belongs to an interval $[\alpha, \beta) \in \mathsf{I}$. Observe that the first inequality in (C.3) is satisfied with equality only when Z is constant over any interval $(a, b) \in \mathsf{I}$. This holds because $\overline{\mathsf{K}}(q) < \mathsf{K}(q)$ at every $q \in (a, b)$. As Y is a solution to (C.2), it must satisfy the inequality in (C.3) with equality, and thus it is constant over (α, β). Then, by its right-continuity, Y is constant over $[\alpha, \beta)$. Note that X is also constant over $[\alpha, \beta)$, by definition. These facts, together with $\int_\kappa^1 \mathsf{X}\ (p)\mathrm{d}p < \int_\kappa^1 \mathsf{Y}(p)\mathrm{d}p$, imply that

$$\int_\alpha^1 \mathsf{Y}(p)\mathrm{d}p > \int_\alpha^1 \mathsf{X}\ (p)\mathrm{d}p = \int_\alpha^1 \mathsf{X}_{\mathrm{PAM}}(p)\mathrm{d}p,$$

which contradicts $\mathsf{Y} \quad {}_{\mathrm{Q}}\mathsf{X}_{\mathrm{PAM}}$.

Thus far we have shown that any right-continuous solution Y to (C.2) satisfies $\mathsf{X} \quad {}_{\mathrm{Q}}\mathsf{Y} \quad {}_{\mathrm{Q}}\mathsf{X}$. Let X , X $\in \mathrm{MPS}(X_{\mathrm{PAM}})$ be so that $X\ (G^{-1}(q)) = \mathsf{X}\ (q)$ and $X\ (G^{-1}(q)) = \mathsf{X}\ (q)$ for all $q \in [0,1)$. (That is, X , X are the transformations of X , X from the quantile space to the type space, respectively.) Then, any right-continuous solution Y to (C.1) satisfies $X \quad Y \quad X$. From the Fan-Lorentz theorem it follows that $E(X\ ) \quad E(Y) \quad E(X\ )$, with the first inequality being

strict when $X \not= Y$, and the second inequality being strict when $X \not= Y$. □

Definition C.2. *The solutions X and X to problem (C.1) characterized by Lemma C.4 are respectively called the* efficiency-minimal *and* efficiency-maximal *solutions to (C.1).*

Lemma C.5. *Let $k : \Theta \to \mathbb{R}$ be bounded and integrable, and let $z_\lambda(\theta) = k(\theta) + \lambda\theta$. Consider*

$$\max_{X \in \mathrm{MPS}(X_{\mathrm{PAM}})} \int_\theta z_\lambda(\theta) X(\theta) dG(\theta). \tag{$\mathbf{Z}_\lambda$}$$

For $\lambda_r > \lambda_l \geq 0$, let X_l, X_r respectively be solutions to $(\mathbf{Z}_{\lambda_l})$ and $(\mathbf{Z}_{\lambda_r})$. Then, $X_l \leq X_r$.

Proof. First, without loss of generality we choose X_l to be an efficiency-maximal solution to $(\mathbf{Z}_{\lambda_l})$. The efficiency-maximal solution exists by Lemma C.4. This is without loss of generality because this efficiency-maximal solution majorizes any other solution to $(\mathbf{Z}_{\lambda_l})$, by Lemma C.4. Similarly, choose X_r to be an efficiency-minimal solution to $(\mathbf{Z}_{\lambda_r})$, without loss of generality.

We next transform $(\mathbf{Z}_\lambda)$ to the quantile space, as in the proof of Lemma C.4. This is done by a change of variables $q = G(\theta)$. Let $\mathsf{z}_\lambda(q) = z_\lambda(G^{-1}(q))$, $\mathsf{X}_l(q) = X_l(G^{-1}(q))$, and $\mathsf{X}_r(q) = X_r(G^{-1}(q))$, for all $q \in [0,1]$. Recall the definitions of $\mathsf{X}_{\mathrm{PAM}}$, $\leq_{\mathrm{Q}}$ and $\mathrm{MPS}_{\mathrm{Q}}$ from the proof of Lemma C.4. The problem $(\mathbf{Z}_\lambda)$ is then transformed to the following problem in the quantile space:

$$\max_{\mathsf{X} \in \mathrm{MPS}_{\mathrm{Q}}(\mathsf{X}_{\mathrm{PAM}})} \int_0^1 \mathsf{z}_\lambda(q)\mathsf{X}(q)\mathrm{d}q. \tag{$\mathbf{Q}_\lambda$}$$

Define $\mathsf{Z}_\lambda(q) = \int_0^q \mathsf{z}_\lambda(\kappa)\mathrm{d}\kappa$. Since X_l, X_r are respectively solutions to $(\mathbf{Z}_{\lambda_l})$ and $(\mathbf{Z}_{\lambda_r})$, then $\mathsf{X}_l, \mathsf{X}_r$ are respectively solutions to $\mathbf{Q}_{\lambda_l}$ and $\mathbf{Q}_{\lambda_r}$.

Let I be the set of all primitive ironed intervals of Z_{λ_l}. Also, let I be the set of all maximal ironed intervals of Z_{λ_r}. Since X_l is an efficiency-maximal solution to $(\mathbf{P}_{\lambda_l})$, then X_l is the extreme point of $\mathsf{X}_{\mathrm{PAM}}$ that is characterized by the family of intervals I. This follows from the proof of Lemma C.4. From the same proof it follows that X_r is the extreme point of $\mathsf{X}_{\mathrm{PAM}}$ that is characterized by the family of intervals I, because X_r is an efficiency-minimal solution to $(\mathbf{P}_{\lambda_r})$. By part (iii) of Lemma C.3, every interval in I is contained in some interval in I. This implies that $\mathsf{X}_l \leq_{\mathrm{Q}} \mathsf{X}_r$. Thus, $X_l \leq X_r$. □

Any point on the Pareto frontier can be generated only by a nonwasteful mechanism. This holds by the same argument that proves the nonwastefulness of optimal mechanisms, Lemma A.5. Based on this observation, we next set up an optimization problem the solutions to which correspond to the points on the Pareto frontier. Let $\lambda \geq 0$, and consider the problem

$$\max_{X \in \mathrm{MPS}(X_{\mathrm{PAM}})} \int_\theta X(\theta)(\lambda\theta + \phi_G(\theta))\mathrm{d}G(\theta). \tag{$\mathbf{P}_\lambda$}$$

When X is the interim allocation rule of a mechanism, recall that $W(X) = \int_\theta X(\theta)\phi_G(\theta)dG(\theta)$ equals the welfare under that mechanism. Similarly, $E(X) = \int_\theta \theta X(\theta)dG(\theta)$ is the efficiency under that mechanism. Thus, the integrant in $(\mathbf{P}_\lambda)$ is $\lambda E(X) + W(X)$, which corresponds to the problem of a planner who assigns different weights to welfare and efficiency. To simplify notation, we let $\lambda = \infty$ correspond to the case where the objective of $(\mathbf{P}_\lambda)$ equals $E(X)$ (where the planner assigns a weight $\lambda_W = 0$ to welfare).

We say λ *generates* (e, w) *2* P *through* X if X is a solution to $(\mathbf{P}_\lambda)$ satisfying $E(X) = e$ and $W(X) = w$. We briefly say that λ generates (e, w) if there exists such a solution X to $(\mathbf{P}_\lambda)$.

Lemma C.6. *Consider* (e, w) *2* P*.*

i. If (e, w) *is an extreme point of the Pareto frontier, then the following holds: there is* $\lambda \quad 0$ *that generates* (e, w) *through a right-continuous extreme point of MPS*(X_{PAM})*. If (not necessarily distinct)* λ_1, λ_2 *generate* (e, w) *respectively through right-continuous* X_1, X_2*, then* $X_1 = X_2$*.*

ii. If (e, w) *is a non-extreme point of the Pareto frontier, then the following holds: There is a maximal open interval* (e_l, e_r) *3* e *over which the Pareto frontier is a ne. Also, there is* $\lambda \quad 0$ *which generates* $(e_l, w\ (e_l))$ *through the e ciency-minimal solution to* $(\mathbf{P}_\lambda)$*, and* $(e_r, w\ (e_r))$ *through the e ciency-maximal solution to* $(\mathbf{P}_\lambda)$*.*

Proof. **Proof of part i.** We first consider the case $(e, w) = (\overline{e}, w\ (\overline{e}))$. This point is generated by the Pareto weight $\lambda = \infty$,[36] because $E(X_{\text{PAM}}) = \overline{e}$ is the highest level of efficiency attainable by a mechanism. Also, since u is strictly supermodular, then $E(X) < E(X_{\text{PAM}})$ for all right-continuous X ~~6=~~ X_{PAM} with $X \quad X_{\text{PAM}}$. This proves the claim when $(e, w) = (\overline{e}, w\ (\overline{e}))$.

We next consider the case where the Pareto frontier is neither left- nor right-affine at e. Then, there is a line passing through (e, w) such that every point of the Pareto frontier is below that line; i.e., there is $\gamma > 0$ such that $\gamma e' + w' \quad \gamma e + w$ for every (e', w') *2* P with the inequality holding strictly for (e', w') ~~6=~~ (e, w). Fix $\lambda = \gamma$ in the problem $(\mathbf{P}_\lambda)$. Every solution X to the problem must satisfy $E(X\ ) = e$ and $W(X\ ) = w$. By Lemma C.4, this can hold only if there is a unique right-continuous solution X_λ to $(\mathbf{P}_\lambda)$. Such a solution must be an extreme point of MPS(X_{PAM}), by Bauer's Maximum Principle. If there is a Pareto weight $\lambda' \neq \lambda$ that generates (e, w) through a right-continuous $X_{\lambda'}$, then $X_{\lambda'}$ must also be a solution to $(\mathbf{P}_\lambda)$. It follows that $X_{\lambda'} = X_\lambda$. This proves part i of the lemma when the Pareto frontier is neither left- nor right-affine at e.

It remains to consider the case where the Pareto frontier is either left- or right-affine at e. Consider the former case: let $e' < e$ be such that the Pareto frontier is affine over the interval $[e', e]$. Observe that there is a unique line l passing through the points $(e, w\ (e))$ and $(e', w\ (e'))$, and that every point on the Pareto frontier lies on or below this line. That is, there is a unique $\gamma > 0$ such that $\gamma e'' + w'' \quad \gamma e + w$ holds for every (e'', w'') *2* P, with the inequality holding with equality

[36] Recall that, by definition, $\lambda = \infty$ corresponds to the case where the objective of $(\mathbf{P}_\lambda)$ equals $E(X)$; i.e., the planner assigns a weight $\lambda_W = 0$ to welfare.

when $(e'', w'') = (e, w)$ or $(e'', w'') = (e', w')$. Fix $\lambda = \gamma$ in the problem $(\mathbf{P}_\lambda)$. Apply Lemma C.4 to this problem to recover the right-continuous extreme point solutions X_λ and X_λ as specified in the lemma statement. It follows that $E(X_\lambda) > E(X)$ for every right-continuous solution $X \neq X_\lambda$ to $(\mathbf{P}_\lambda)$. As the Pareto frontier is left-affine at (e, w), then λ generates (e, w) through X_λ. If there is a Pareto weight $\lambda' \neq \lambda$ that generates (e, w) through some right-continuous $X_{\lambda'}$, then $X_{\lambda'}$ must also be a solution to $(\mathbf{P}_\lambda)$. It follows that $X_{\lambda'} = X_\lambda$. This proves the lemma when the Pareto frontier is left-affine at (e, w). The proof for the right-affine case is identical, but for X_λ replaced with X_λ and the inequality $E(X_\lambda) > E(X)$ replaced with $E(X_\lambda) < E(X)$ in the above argument.

Proof of part ii. As the function w (which characterizes the Pareto frontier P) is concave and decreasing, then a point $(e, w) \in P$ can be written as a convex combination of two other points on the Pareto frontier if and only if (e, w) is an interior point of an affine line segment that lies on the Pareto frontier. We define $(e_l, w(e_l))$ and $(e_r, w(e_r))$ to be the endpoints of this line segment. The line segment has a negative slope; let $\lambda > 0$ denote the magnitude of the slope.

Observe that the Pareto frontier is left-affine at e_r. Thus, the argument in part i for the case where e is left-affine applies here with $e = e_r$. From that argument it follows that λ generates $(e_r, w(e_r))$ through X_λ, which is an efficiency-maximal solution to $(\mathbf{P}_\lambda)$. Similarly, we observe that the Pareto frontier is right-affine at e_l. Then, λ generates $(e_l, w(e_l))$ through X_λ, which is an efficiency-minimal solution to $(\mathbf{P}_\lambda)$, by definition. The proof is complete. □

Proof of Theorem C.1, Part i. Consider an extreme point (e, w) of the Pareto frontier. By Lemma C.6, there is a $\lambda > 0$ that generates (e, w) through a right-continuous extreme point X_λ of $\text{MPS}(X_{\text{PAM}})$. Moreover, by the same lemma, there is no right-continuous $X \in \text{MPS}(X_{\text{PAM}})$ which satisfies $E(X) = e$, $W(X) = w$, and $X \neq X_\lambda$. Define $\mu(e)$ to be the disclosure policy $\mu[X_\lambda]$. By Lemma C.1, $\mu[X_\lambda]$ is the unique disclosure policy that implements X_λ, up to a relabeling of the realization space. By the same lemma, X_λ also pools adjacent object types. This proves part i.

Proof of Theorem C.1, Part ii. There is $\lambda > 0$ which generates (e_l, w_l) through a right-continuous extreme point $X_l \in \text{MPS}(X_{\text{PAM}})$, and generates (e_r, w_r) through a right-continuous extreme point $X_r \in \text{MPS}(X_{\text{PAM}})$. Moreover, X_l, X_r are respectively the efficiency-minimal and efficiency-maximal solutions to $(\mathbf{P}_\lambda)$ The two latter facts hold by part ii of Lemma C.6. By Lemma C.4, X_r majorizes X_l. Thus, $\mu[X_l]$ is weakly less informative than $\mu[X_r]$, by Lemma C.2. Denote the two latter policies by μ_l, μ_r and their realization spaces by $\widehat{\Omega}_l, \widehat{\Omega}_r$, respectively. To construct $\mu(e')$, let μ_l, μ_r be disclosure policies respectively equal to μ_l, μ_r, with the difference that the realization spaces of μ_l, μ_r are relabeled so that they contain no common element.

Let $\delta \in (0, 1)$ be such that $e' = \delta e_l + (1 - \delta)e_r$. Define $\mu(e')$ to be the disclosure policy that sends a signal realization using $\mu(e_l)$ with probability δ, and otherwise sends a signal realization using $\mu(e_r)$. It follows that $\mu(e')$ implements $\delta X_l + (1 - \delta)X_r$. Denote the last expression by Z. Hence $\mu(e')$ generates the point $(E(Z), W(Z))$. By the linearity of the functionals E, W,

$E(Z) = \delta E(X_l) + (1 \quad \delta)E(X_r) = e^0$ and $W(Z) = \delta W(X_l) + (1 \quad \delta)W(X_r) = w^0$. Thus, $\mu(e^0)$ generates (e^0, w^0). Moreover, since X_r majorizes X_l, then X_r majorizes Z, and Z majorizes X_l. Hence, $\mu[Z]$ is less informative than $\mu[e_r]$ and more informative than $\mu[e_l]$. This proves part ii.

Proof of Theorem C.1, Part iii. For the proof we will consider two points (e, w) and (e^0, w^0) on the Pareto frontier with $e < e^0$. We will consider two separate cases and show that (1) if the function w (which defines the Pareto frontier) is affine over the interval (e, e^0), then $\mu(e^0)$ is more informative than $\mu(e)$, and (2) the same conclusion holds if the Pareto frontier is not affine over the interval (e, e^0). These two facts together prove the claim.

We first consider case (1). Let $[e_l, e_r]$ be the maximal closed interval containing e, e^0 such that w is affine over (e_l, e_r). In part ii of the theorem we showed that, for e^{00} 2 $[e_l, e_r]$, the policy $\mu(e^{00})$ is more informative for higher values of e^{00}.

It remains to consider case (2), where the Pareto frontier is not affine over (e, e^0). In this case, we define e_l, e_r as follows. If the Pareto frontier is right-affine at e, then let e_l be the largest real number such that the Pareto frontier is affine over the interval (e, e_l). Otherwise, $e_l = e$. If the Pareto frontier is left-affine at e^0, then let e_r be the smallest real number such that the Pareto frontier is affine over the interval (e_r, e^0). Otherwise, $e_r = e^0$. Define $w_l = w$ (e_l) and $w_r = w$ (e_r). Observe that $e_l \quad e_r$, and that (e_l, w_l) and (e_r, w_r) are extreme points of the Pareto frontier.

To prove that $\mu(e^0)$ is more informative than $\mu(e)$, it suffices to show that $\mu(e_r)$ is more informative than $\mu(e_l)$. This is due to part ii of the theorem, which we proved above. If the function w is affine over the interval (e_l, e_r), then the claim follows from case (1). Otherwise, there are $\lambda_l, \lambda_r \quad 0$ that generate e_l, e_r respectively. This is due to part i of Lemma C.6. Since $e_r > e_l$, then $\lambda_r > \lambda_l$. Let X_l, X_r respectively denote the extreme points of $\text{MPS}(X_{\text{PAM}})$ through which λ_l, λ_r generate the points $(e_l, w_l), (e_r, w_r)$. That also means X_l, X_r are solutions to the problems $(\mathbf{P}_{\lambda_l}), (\mathbf{P}_{\lambda_r})$. Then, X_r majorizes X_l, by Lemma C.5. Thus, $\mu[X_r]$ is more informative than $\mu[X_l]$, by Lemma C.2. By part i of the theorem, $\mu[X_l]$ and $\mu[X_r]$ respectively equal $\mu(e_l)$ and $\mu(e_r)$ up to relabeling the realization spaces. Hence, $\mu(e_r)$ is more informative than $\mu(e_l)$.

C.3 Proof of Theorem 5.1

Fix λ_E and consider $\lambda_W, \lambda_W^0 \quad 0$ with $\lambda_W < \lambda_W^0$. Let μ, μ^0 be disclosure policies that attain the second-best objective value when the planner's weight on welfare is λ_W and λ_W^0, respectively. Let (e, w) and (e^0, w^0) be points on the Pareto frontier generated by μ, μ^0. Hence, $e \quad e^0$. If $e = e^0$, then $(e, w) = (e^0, w^0)$ is an extreme point of the Pareto frontier. Thus, by part (i) of Theorem C.1, μ equals μ^0 up to a relabeling of the realization space, which proves the claim.

Thus, suppose that $e > e^0$. The proof then considers two cases, as in the proof of part (iii) of Theorem C.1: Either the Pareto frontier w ( ) is affine over (e^0, e), or not. In the former case, at least one of $(e, w), (e^0, w^0)$ should be an extreme point of the Pareto frontier, since λ_W 6= λ_W^0. Then, from parts (i) and (ii) of Theorem C.1 it follows that μ is more informative than μ^0. It remains to

consider the case where w is *not* affine over $(e^{\emptyset}, e)$. This is precisely case (2) in the proof of part (iii) of Theorem C.1. The same argument applies and implies that μ is more informative than $\mu^{\emptyset}$.

Online Appendix

I Uniqueness properties of monotone disjoint queue mechanisms

We first focus on the allocation of objects to queues, and then on the allocation of the objects sent to each queue to the agents within that queue.

I.1 Allocation of objects across queues

The next theorem is the main result of this subsection. It shows that any mechanism that generates an extreme point of the Pareto frontier has an interim allocation rule X that is an extreme point of $\text{MPS}(X_{\text{PAM}})$. We recall that such X is characterized by a family of disjoint intervals, due to Theorem 0, and can be implemented by a monotone disjoint queue mechanisms, by Theorem 3.1. Moreover, X can be implemented by only pooling adjacent object types, as shown in part (i) of Theorem C.1. This, e.g., implies that any disjoint queue mechanism that implements X must be a *monotone* disjoint queue mechanism.

For the theorem, we recall the definitions of efficiency $E(X) = \int_0^{\bar{\theta}} u(\theta, X(\theta))\,\mathrm{d}G(\theta)$ and welfare $W(X) = \int_0^{\bar{\theta}} \phi(\theta, X(\theta))\,\mathrm{d}G(\theta)$ for a given X 2 $\text{MPS}(X_{\text{PAM}})$.

Theorem I.1. *Consider an extreme point (e, w) of the e ciency-welfare Pareto frontier and X 2 MPS(X_{PAM}) with $E(X) = e$ and $W(X) = w$. Then, X is an extreme point of MPS(X_{PAM}).*

Proof. Let $X = \text{MPS}(X_{\text{PAM}})$ and $\boldsymbol{\lambda} = (\lambda_E, \lambda_W)$. Define the functional $\Phi_{\boldsymbol{\lambda}} : \mathsf{L}_1 \;!\; \mathsf{R}_+$ by $\Phi_{\boldsymbol{\lambda}}(X) = \lambda_E E(X) + \lambda_W W(X)$. Also, let $\mathsf{OPT}_{\boldsymbol{\lambda}} = \max_{X 2 X} \Phi_{\boldsymbol{\lambda}}(X)$. Define $\overline{X} = X_{\text{PAM}}$, and let $\underline{X}$ be the interim allocation rule that equals $\int_0^{\bar{\theta}} \overline{X}(\theta)\,\mathrm{d}G(\theta)$ everywhere in its domain.

Let (e, w) be an arbitrary extreme point of the Pareto frontier. Choose $\boldsymbol{\lambda} = (\lambda_E, \lambda_W)$ such that $e^{\emptyset}\lambda_E + w^{\emptyset}\lambda_W \quad e\lambda_E + w\lambda_W$ for every point $(e^{\emptyset}, w^{\emptyset})$ on the Pareto frontier. (This is possible because the Pareto frontier is concave and decreasing, by Fact C.1.) There is $X \;\; 2 X$ such that $\Phi_{\boldsymbol{\lambda}}(X) = \mathsf{OPT}_{\boldsymbol{\lambda}}$. Consider the inner product space defined over L_1 with the inner product operator $\langle X, Y \rangle = \int_0^{\bar{\theta}} X(\theta)Y(\theta)\,\mathrm{d}G(\theta)$. Since $\Phi_{\boldsymbol{\lambda}}$ is convex by Lemma A.3, and also lower semicontinuous, then it is subdifferentiable at X (Gretsky et al., 2002).[1] Let G be a subgradient of $\Phi_{\boldsymbol{\lambda}}$ at X. Thus,

$$\Phi_{\boldsymbol{\lambda}}(X) \quad \Phi_{\boldsymbol{\lambda}}(X) + \langle G, X \quad X \rangle.$$

Denote the right-hand side by $\Psi_{\boldsymbol{\lambda}}(X)$. For a functional $\Gamma : \mathsf{L}_1 \;!\; \mathsf{R}_+$ we say X 2 L_1 is a *maximizer of Γ over X* if $\Gamma(X) = \max_{Y 2 X} \Gamma(Y)$.

Claim I.1. *Any maximizer of $\Psi_{\boldsymbol{\lambda}}$ over X is also a maximizer of $\Phi_{\boldsymbol{\lambda}}$ over X.*

[1] See (Gretsky et al., 2002) for generic sufficient conditions for subdifferentiability of convex functionals. In particular, convexity and lower semicontinuity imply that a convex functional is subdifferentiable at every interior point of its domain.

Proof. We then observe that

$$\Psi_{\boldsymbol{\lambda}}(X) \quad \Phi_{\boldsymbol{\lambda}}(X) \quad \Phi_{\boldsymbol{\lambda}}(X') = \Psi_{\boldsymbol{\lambda}}(X') = \mathsf{OPT}_{\boldsymbol{\lambda}},$$

where the first and last inequalities hold by the definition of subgradient, and the second one holds by X being a maximizer of $\Phi_{\boldsymbol{\lambda}}$. If X maximizes $\Psi_{\boldsymbol{\lambda}}$ over $\mathcal{X}$, then $\Psi_{\boldsymbol{\lambda}}(X) = \mathsf{OPT}_{\boldsymbol{\lambda}}$, from which it follows that X also maximizes $\Phi_{\boldsymbol{\lambda}}$ over $\mathcal{X}$. □

Let $\mathcal{Y}$ be the set of maximizers of $\Psi_{\boldsymbol{\lambda}}$ over $\mathcal{X}$. (For brevity, henceforth we drop the phrase "over $\mathcal{X}$" from this expression throughout the rest of the proof.) By Lemma C.4, there exist $Y', Y'' \in \mathcal{Y}$ such that Y', Y'' are extreme points of $\mathcal{X}$ and

$$Y \quad Y' \quad Y'', \text{ for all } Y \in \mathcal{Y}. \tag{I.1}$$

We now consider two cases to complete the proof: The first case assumes that the Pareto frontier is neither right-affine nor left-affine at e, and the second case assumes that the Pareto frontier is either right-affine or left-affine at e.

To prove the claim in the first case, recall that any maximizer of $\Psi_{\boldsymbol{\lambda}}$ is also a maximizer of $\Phi_{\boldsymbol{\lambda}}$, by Claim I.1. In particular, X', Y', Y'' are (not necessarily distinct) maximizers of $\Phi_{\boldsymbol{\lambda}}$. By the Fan-Lorentz theorem, $E(Y') \quad E(X') \quad E(Y'')$. Then, since (e, w) is an extreme point of the Pareto frontier, and since the Pareto frontier is neither right- nor left-affine at (e, w), then $E(Y') = E(X') = E(Y'')$. Thus, $Y' = Y''$; i.e., the set $\mathcal{Y}$ is singleton and contains X'. This is due to the Fan-Lorentz theorem and strict supermodularity of u. The claim is then proved since $X' = Y'$ and Y' is an extreme point of $\mathcal{X}$.

It remains to prove the claim in the second case; i.e., the Pareto frontier is either right- or left-affine at e. We prove the claim assuming right-affinity, the proof for the left-affine case is symmetric. Suppose that the Pareto frontier is affine over the interval $[e, \bar{e}]$ for $\bar{e} > e$. Let $\bar{w} \in \mathbb{R}_+$ such that $(\bar{e}, \bar{w})$ is on the Pareto frontier. Thus far, we have chosen $\boldsymbol{\lambda} \in \mathbb{R}^2_+$ so that $e^0\lambda_E + w^0\lambda_W \quad e\lambda_E + w\lambda_W$ for every point (e^0, w^0) on the Pareto frontier. For the rest of the argument, we choose $\boldsymbol{\lambda}$ such that, in addition to the previous condition, $\bar{e}\lambda_E + \bar{w}\lambda_W = e\lambda_E + w\lambda_W$ is satisfied. This is possible since every point on the line segment connecting (e, w) and $(\bar{e}, \bar{w})$ belongs to the Pareto frontier.

Let $\mathcal{Y}$ be the set of maximizers of $\Psi_{\boldsymbol{\lambda}}$. By Lemma C.4, there are $Y', Y'' \in \mathcal{Y}$ that are extreme points of $\mathcal{X}$ and satisfy (I.1). By (I.1), $E(Y') \quad E(X') \quad E(Y'')$. Since (e, w) is an extreme point of the Pareto frontier, and since the Pareto frontier is right-affine at (e, w), then $E(Y') = E(X')$. Moreover, $E(Y') < E(Y)$ if $Y \in \mathcal{Y}$ and $Y \neq Y'$, due to the Fan-Lorentz theorem and strict supermodularity of u. Thus, $X' = Y'$. The proof concludes since Y' is an extreme point of $\mathcal{X}$. □

A corollary of the above theorem is that any disjoint queue mechanism that generates an extreme

point of the Pareto frontier must be a *monotone* disjoint queue mechanism.

I.2 Allocation of objects within queues

The FIFO queuing scheme is used in monotone disjoint queue mechanisms for allocating objects to agents within each queue. We define the notion of queuing scheme formally below. Loosely speaking, a queuing scheme is a probability distribution over the queue positions according to which an arriving object is allocated to one of the positions. We will show that FIFO is a *universally optimal* queuing scheme, meaning that, in any market, an optimal mechanism that uses schemes other than FIFO can instead adopt FIFO while preserving optimality. We also show that FIFO is the unique scheme that satisfies this property. In Online Appendix I.3 we prove the same result but for an alternative definition that defines a queuing scheme as a probability distribution over positions that is *scaled* according to the queue length. SIRO is an example of such schemes.

The *position* of an agent a at a time t in a queue is defined as the measure of agents present in that queue at time t who have arrived before a divided by the objects' arrival rate to the queue.[2] (The normalization simplifies notation but is inconsequential, since the arrival rate of objects sent to the queue is held fixed throughout the proof.) A position p is *better* than p' if $p < p'$.

A queuing *scheme* is a tuple $(\mathsf{s}, \bar{p})$ where $\bar{p} \quad 0$ and s is a probability measure over the interval $[0, \bar{p}]$ and this interval is the minimal closed interval containing the support of s. The probability that an agent who joins the queue is matched at a position weakly better than p is $\int_0^p \mathrm{d}\mathsf{s}(t)$. The minimality assumption captures the fact that positions in a queue exist only if agents that reach those positions are eventually matched (possibly later at better positions). We sometimes simply use s to refer to the queuing scheme when there is no risk of confusion regarding its support. With slight abuse of notation, we use $\mathsf{s}(x)$ to denote $\int_0^x \mathrm{d}\mathsf{s}(y)$.

A *time index* is a decreasing function $\mathsf{p} : \mathsf{R}_+ \; ! \; \mathsf{R}_+$ that is not zero everywhere and $\lim_{t \,!\, 1} \mathsf{p}(t) = 0$. The interpretation is that $\mathsf{p}(t)$ gives the position of an agent with waiting time t, and that there is no positive mass of agents in the queue who wait for an infinite amount of time yet are never matched. (This cannot happen as every agent who joins the queue should be eventually matched.)

Definition I.1. *A queuing scheme* $(\mathsf{s}, \bar{p})$ *is* viable *if there is a time index* p *with* $\mathsf{p}'(t) = \int_0^{\mathsf{p}(t)} \mathrm{d}\mathsf{s}(x)$ *when* $\mathsf{p}(t) > 0$*. When this condition holds, we say the time index* p *is* consistent *with the scheme* s*.*

The condition says that the rate by which the position of an agent decreases (i.e., gets better) at every time equals the rate of agents ahead of her who get matched at that time.

Proposition I.1. *Any queuing scheme* $(\mathsf{s}, \bar{p})$ *is viable. If* $1/\mathsf{s}(x)$ *is integrable over* $[0, \bar{p}]$*, then the following holds: For every* $p_0 \quad \bar{p}$ *there is a unique time index* p *consistent with* s *such that*

[2] The objects' rate of arrival equals the agents' rate of arrival to the same queue by the steady-state condition.

$\mathsf{p}(0) = p_0$.[3] *Moreover,* p *is characterized by*

$$\mathsf{p}(t) = p_0 \quad t \text{ if } t \quad \Delta, \tag{I.2}$$

$$\int_0^{\mathsf{p}(t)} \frac{1}{\mathsf{s}(x)}\,\mathrm{d}x = \bar{t} \quad t + \Delta \text{ if } \Delta < t \quad \Delta + \bar{t}, \tag{I.3}$$

$$\mathsf{p}(t) = 0 \text{ if } t > \Delta + \bar{t}, \tag{I.4}$$

where $\Delta = p_0 \quad \bar{p}$ *and* $\bar{t} = \int_0^{\bar{p}} \frac{1}{\mathsf{s}(x)}\,\mathrm{d}x$. *Speci cally,* p *is continuous and strictly increasing over* $[0, p_0]$.

On the other hand, if $\frac{1}{\mathsf{s}(x)}$ *is not integrable over* $[0, \bar{p}]$, *then* $\lim_{\epsilon\,!\,0} \int_\epsilon^{\bar{p}} \frac{1}{\mathsf{s}(x)}\,\mathrm{d}x = 1$, *and the following holds: For every* $p_0 \quad \bar{p}$ *there is a unique time index* p *consistent with* s *such that* $\mathsf{p}(0) = p_0$. *Moreover,* p *is characterized by*

$$\mathsf{p}(t) = p_0 \quad t \text{ if } t \quad \Delta,$$
$$\int_{\mathsf{p}(t)}^{\bar{p}} \frac{1}{\mathsf{s}(u)}\,\mathrm{d}u = t \quad \Delta \text{ if } t > \Delta.$$

The intuition for this proposition is that $\frac{1}{\mathsf{s}(x)}$ is the reciprocal of the rate of change in position, i.e., the amount of time spent at each position. Integrating the reciprocal over all positions thus gives the total time $\bar{t}$ that it takes for an agent to get from position $\bar{p}$ to position 0. For example, the FIFO queuing scheme is the Dirac measure at 0, which we denote by s_F. For every $p_0 \quad 0$, there is a unique time index consistent with s_F, which satisfies $\mathsf{p}'(t) = \quad 1$ for all t *2* $(0, p_0]$. That is, the position in the queue at all times decreases at a rate equal to the objects' arrival rate.

Proof of Proposition I.1. We start with proving the claim for when the integrability condition holds. For a position p *2* $(\bar{p}, p_0]$, we have $\mathsf{s}(p) = 1$, which means that any time index p consistent with s must satisfy $\mathsf{p}'(t) = \quad 1$ when $\mathsf{p}(t) > \bar{p}$. Hence, an agent at position p_0 (with waiting time 0) gets to position $p \quad \bar{p}$ after waiting for $p_0 \quad p$ units of time. Thus, any p consistent with s satisfies (I.2).

Next, we compute the time that it takes for an agent at position $\bar{p}$ to get to position 0 (assuming that the agent is not matched before reaching position 0). Denote this time by $\bar{t}$. (That is, $\bar{t}$ is such that $\mathsf{p}(\bar{t} + \Delta) = 0$.) Observe that $\bar{t}$ must satisfy the following:

$$\frac{\mathsf{p}'(t)}{\int_0^{\mathsf{p}(t)} \mathrm{d}\mathsf{s}(x)} = \frac{\mathsf{p}'(t)}{\mathsf{s}(\mathsf{p}(t))} = \quad 1$$
$$, \quad \int_\Delta^{\bar{t}+\Delta} \frac{\mathsf{p}'(t)}{\mathsf{s}(\mathsf{p}(t))}\,\mathrm{d}t = \quad \bar{t}$$
$$\text{(change of variable } u = \mathsf{p}(t)) \quad , \quad \int_0^{\bar{p}} \frac{1}{\mathsf{s}(u)}\,\mathrm{d}u = \bar{t}. \tag{I.5}$$

[3] Thus, the choice of p depends on $\bar{p}$; for simplicity, we are dropping a subscript $\bar{p}$ from the notation for p.

Similarly, we consider an agent with waiting time $t \in (\Delta, \bar{t}+\Delta)$ and characterize the position associated with such a waiting time as follows:

$$\int_t^{\bar{t}+\Delta} \frac{\mathsf{p}'(t)}{\mathsf{s}(\mathsf{p}(t))}\,\mathrm{d}t = \quad (\bar{t}+\Delta \quad t)$$
$$\text{(change of variable } u=\mathsf{p}(t)) \;,\; \int_0^{\mathsf{p}(t)} \frac{1}{\mathsf{s}(u)}\,\mathrm{d}u = \bar{t}+\Delta \quad t, \text{ for } t \in (\Delta, \bar{t}+\Delta).$$

Every time index p consistent with s has to satisfy the above equation. Define $\mathsf{f}(z) = \int_0^z \frac{1}{\mathsf{s}(u)}\,\mathrm{d}u$. For any $t \in (\Delta, \bar{t}+\Delta)$, the right-hand side of the above displayed equation is inside $(0, \bar{t})$. Since f is a continuous and strictly increasing function, and $\mathsf{f}(0)=0$ and $\mathsf{f}(\bar{p})=\bar{t}$, then there is a unique $z_t \in (0,\bar{p})$ such that $\mathsf{f}(z_t) = \bar{t}+\Delta \quad t$. This means that $\mathsf{p}(t) = z_t$. Thus, the value of $\mathsf{p}(t)$ is uniquely pinned down as well when $t > \Delta$, and (I.3) holds. Finally, (I.4) holds because p is decreasing and $\mathsf{p}(\bar{t}+\Delta)=0$. This proves the claim when the integrability condition holds. The fact that $\mathsf{p}(t)$ is continuous and strictly increasing follows from the fact that f satisfies these properties.

Now, suppose that $1/\mathsf{s}(x)$ is not integrable over $[0,\bar{p}]$. As $1/\mathsf{s}(x)$ is decreasing, then $\mathsf{s}(0)=0$ and $\lim_{\epsilon \to 0} \int_\epsilon^{\bar{p}} \frac{1}{\mathsf{s}(x)}\,\mathrm{d}x = \infty$ must hold for nonintegrability to hold. For any time index p consistent with s and $t \quad 0$, similar to (I.5) we write

$$\frac{\mathsf{p}'(t)}{\int_0^{\mathsf{p}(t)} \mathrm{d}\mathsf{s}(x)} = \frac{\mathsf{p}'(t)}{\mathsf{s}(\mathsf{p}(t))} = \quad 1$$
$$\;,\; \int_\Delta^{t+\Delta} \frac{\mathsf{p}'(t)}{\mathsf{s}(\mathsf{p}(t))}\,\mathrm{d}t = \quad t$$
$$\text{(change of variable } u=\mathsf{p}(t)) \;,\; \int_{\mathsf{p}(t+\Delta)}^{\bar{p}} \frac{1}{\mathsf{s}(u)}\,\mathrm{d}u = t.$$

Define $\mathsf{f}(z) = \int_z^{\bar{p}} \frac{1}{\mathsf{s}(u)}\,\mathrm{d}u$. Since f is a strictly increasing function, and $\mathsf{f}(0) = \infty$ and $\mathsf{f}(\bar{p})=0$, then there is a unique $z_t \in (0,\bar{p})$ such that $\mathsf{f}(z_t)=t$. Therefore, $\mathsf{p}(t)=z_t$. Thus, the value of $\mathsf{p}(t)$ is uniquely pinned down when $t > \Delta$. For $t \quad \Delta$, the same argument as in the previous case (where the integrability condition holds) pins down the value of $\mathsf{p}(t)$. □

Corollary I.1. *Let $(\mathsf{s},\bar{p})$ be a queuing scheme and p be a time index consistent with it. Then, p is invertible; i.e., the function p^{-1} exists.*

Proof. The invertibility follows from p being continuous and strictly increasing, which holds by Proposition I.1. □

Using the above corollary, one can characterize the time that it takes for an agent to reach a position x in a queue that she has joined (conditional on the agent not being matched before reaching that position). Suppose that the agent joins a queue with the queuing scheme $(\mathsf{s},\bar{p})$, and let $p_0 > 0$ be the position at which the agent starts waiting immediately after arrival. By

Proposition I.1, there is a unique time index p consistent with the queuing scheme $(\mathsf{s}, \bar{p})$ such that $\mathsf{p}(0) = p_0$. The expected waiting cost that the agent incurs in the queue then equals

$$\int_0^{\mathsf{p}(0)} c(\mathsf{p}^{-1}(x))\, \mathrm{d}\mathsf{s}(x). \tag{I.6}$$

The term inside the integral is the cost incurred from being matched at a position x, which is integrated according to the probability measure over the position of getting matched.

Indirect Queue-Based Mechanisms An *indirect queue-based* mechanism[4] N is defined by a finite number of queues $q_1, \ldots, q_k$ where objects of type $\omega \in \Omega$ are sent to each q_i at a rate of $r_{\omega,i}$, and q_i is associated with (i) a queuing scheme $(\mathsf{s}_i, \bar{p}_i)$ and (ii) a family of distributions $\{F_i(p)\}_{p \in [0,\bar{p}_i]}$. Each distribution $F_i(p)$ is over the object types Ω, and its probability mass function is denoted by $F_i(p, \cdot) : \Omega \to \mathbb{R}_+$. For every $\omega \in \Omega$, the function $F_i(\cdot, \omega)$ is integrable with respect to s_i. If an agent is matched by the queuing scheme s_i when she is at a position p, then the object type allocated to her has the distribution $F_i(p)$. The option of departing the market immediately after arrival corresponds to joining a queue q_0 where agents are immediately allocated an object of type ω_0.

For the next definition, let $K = \{1, \ldots, k\}$. Also, with slight abuse of notation let $G(S)$ denote the measure assigned by G to a measurable subset $S \subseteq \Theta$.

Definition I.2. *A* steady state *of the indirect queue-based mechanism N is defined by a set of time indices $\{\mathsf{p}_i\}_{i \in K}$ and a partition of the set of agent types to k measurable subsets $\{\Theta_i\}_{i \in K}$, such that*

(i) the queuing schemes are executed: *For every $i \in K$, p_i is consistent with $(\mathsf{s}_i, \bar{p}_i)$;*

(ii) in each queue, demand equals supply*: For $i \in K$ and $\omega \in \Omega$, $G(\Theta_i) \int_0^{\mathsf{p}_i(0)} F_i(p, \omega)\, \mathrm{d}\mathsf{s}_i(p) = r_{\omega,i}$; and*

(iii) incentive compatibility *holds: the agents in Θ_i attain a higher expected payoff from joining q_i than from joining any other queue, where an agent's expected payoff from joining q_i is the expected utility from joining that queue (computed according to s_i) minus the expected waiting cost (computed according to p_i).*[5]

In the above definition, we say that p_i is the time index *associated* with queue q_i. The following lemma shows that the planner attains the same objective value in all of the steady states of an indirect queue-based mechanism.

Lemma I.1. *The payoff attained by an agent of type θ is the same across all steady states of a queue-based mechanism, for all but finitely many $\theta \in \Theta$. The same holds for the utility attained by an agent of type θ.*

[4] As it becomes clear shortly, the mechanism is called indirect since the agents choose which queue to join, as opposed to being assigned to a queue by the mechanism.

[5] We note that the expected utility is well-defined by the integrability assumption, and the expected waiting cost is given by (I.6).

Proof. Consider the (arbitrary) indirect queue-based mechanism N defined above and its steady state defined in Definition I.2. For every queue q_i, let $\bar{c}_i, \bar{\omega}_i$ respectively be the expected waiting cost incurred from joining q_i and the expected object type sent to q_i in that steady state. Without loss of generality suppose that $\omega_0 = \bar{w}_0 \quad \bar{\omega}_1 \quad \dots \quad \bar{\omega}_k$. (Recall that ω_0 is the only object type sent to q_0.) For $i, j \in K$ with $i \quad j$, we must have $\bar{c}_i \quad \bar{c}_j$. Otherwise, no agent will join q_i, which would contradict Condition (ii) in Definition I.2.

Since the agents' utility function u is supermodular, there are no two agents such that the higher-type agent joins a lower-indexed queue. It follows that there exist $0 = \theta_0, \dots, \theta_{k+1} = \overline{\theta}$ such that agents of a type belonging to (θ_i, θ_{i+1}) join q_i, for all i. Then, the incentive-compatibility condition and the envelope condition (Lemma A.1) together uniquely determine the expected waiting cost at each queue. Hence, for all i, the expected object type allocated to an agent of type $\theta \in (\theta_i, \theta_{i+1})$ is the same across all steady states. Also, her expected waiting cost is the same across all steady states. This concludes the proof. □

An indirect queue-based mechanism N is *optimal* if the planner's objective value in some steady state of N equals the second-best objective value (which is her objective value in the optimal mechanism, such as the optimal monotone disjoint queue mechanism of Theorem 3.1). By Lemma I.1, if this condition holds at some steady state of N, it holds at all of its steady states.

An indirect queue-based mechanism N is *detail-free* if (i) it uses the same queuing scheme in all queues with a positive index and (ii) in each such queue the distribution of object types allocated to each position is identical across all positions: $F_{i,p} = F_{i,p'}$ for all $p, p' \in [0, \bar{p}_i]$.

Given an indirect queue-based mechanism N and a queuing scheme s, define the *s-alteration of N*, denoted by $N[\mathsf{s}]$, as follows: $N[\mathsf{s}]$ is a detail-free mechanism involving queues $q_0', \dots, q_k'$ where (i) for every q_i', the queuing scheme used in q_i' is s, and (ii) the distribution of object types sent to q_i' is the same as the distribution of object types sent to q_i. Intuitively, $N[\mathsf{s}]$ sends objects to queues at the same rate as in N; however, it uses the queuing scheme s to allocate objects within each queue, and is agnostic to the object types arriving at that queue.

Definition I.3. *A queuing scheme s is* universally optimal *if for every market M and an optimal queue-based mechanism N for that market, $N[\mathsf{s}]$ is also an optimal mechanism for market M.*

Proposition I.2. *The FIFO queuing scheme is the unique universally optimal queuing scheme.*

Proof. First, we show that FIFO is universally optimal. Consider an arbitrary market M and an optimal queue-based mechanism N in that market involving queues $q_0, \dots, q_k$. Let $\bar{w}_i$ denote the expected object type sent to q_i and let $\bar{c}_i$ denote the expected waiting cost that an agent incurs in queue i. We choose M such that $\bar{c}_i > 0$ for $i \quad 1$.

Let $N' = N[\mathsf{s}_F]$ (i.e., N' is the same as N but uses the FIFO queuing scheme in all queues). Observe that N' has at least one steady state, where the queue lengths are such that the waiting cost at q_i equals $\bar{c}_i$ and every agent joins the same queue in N' and N. This is a valid steady state:

Conditions (i) and (ii) in Definition I.2 hold by definition. Condition (iii) holds because for every agent and any q_i, the agent's expected utility and expected waiting cost from joining q_i is the same in N and in the constructed steady state for N , by construction. Thus, the planner's objective value is the same in N and in the constructed steady state for N . From Lemma I.1 it follows that the latter statement holds for any steady state of N . That is, N is optimal.

The proof for uniqueness of FIFO is by contradiction. Let $(\mathsf{s}, \bar{p})$ be a universally optimal queuing scheme that is not FIFO. Then, $\bar{p} > 0$, and there is $\delta > 0$ and p 2 $(0, \bar{p})$ such that $\mathsf{s}(p\ ) > \delta$. By the proof of Proposition I.1, if s is used in an indirect queue-based mechanism, then $t = \int_p^{\bar{p}} \frac{1}{\mathsf{s}(x)}\, \mathrm{d}x$ is the time that it takes for an agent to get from position $\bar{p}$ to position p in that mechanism. Let M be an arbitrary market with $\Omega = f\omega_0, \ldots, \omega_n g$ and a waiting cost function c such that $c(t\ ) > u(\bar{\theta}, \omega_n)/\delta$. Let N be any optimal indirect queue-based mechanism for M (e.g., N can be based on a monotone disjoint queue mechanism, which always exists by Theorem 3.1). Let $N = N[\mathsf{s}]$. Consider an arbitrary steady state of N where every queue q_i of N is associated with a time index p_i.

Recall that $[0, \bar{p}]$ is the minimal interval containing the support of s. From Condition (ii) in Definition I.2 it follows that $\mathsf{p}_i(0)$ $\bar{p}$ in every queue q_i with i 1. Since any time index consistent with s is continuous (Proposition I.1), then at any steady state of N an agent incurs a waiting cost no smaller than $c(t\ )$ with probability at least δ. That means the expected payoff from joining the queue is at most $u(\bar{\theta}, \omega_n)$ $\delta c(t\ ) < 0$, This, however, contradicts individual rationality, because agents can obtain a nonnegative payoff from joining q_0. □

I.3 Queuing schemes that scale according to queue length

In practice, sometimes a policy refers to a family of queuing schemes that are obtained from *scaling* one another. For example, the SIRO policy corresponds to a family of queuing schemes $f\mathsf{s}_p g_{p>0}$ where s_p is the uniform measure over the interval $[0, p]$ and p denotes the measure of agents present in the queue at the steady state normalized by the agents' arrival rate. We next formally define the class of such policies and show that none of them can be universally optimal. Intuitively, the reason is that under these policies the payoff that an agent achieves from remaining in the queue can decrease over time and become negative, which leads to the failure of individual rationality.

A *normalized queuing scheme* is a probability measure s over the unit interval, such that this interval is the minimal interval containing the support of s. Let $\mathsf{s}(x) = \int_0^x \mathrm{d}\mathsf{s}(y)$. For p 2 R_+, define the queuing scheme (s_p, p) by $\mathsf{s}_p(q) = \mathsf{s}(q/p)$ for all q 2 $[0, p]$. Call s_p a *scaled version of* s.

Let N be an indirect queue-based mechanism involving queues $q_0, \ldots, q_k$, and s be a normalized queuing scheme. An *s-alternation of N* is a (direct) mechanism involving queues $q_0^{\emptyset}, \ldots, q_k^{\emptyset}$ in which the allocation of objects and agents to queues is defined as follows

(i) *the allocation of objects to queues is the same as in N:* each object type is sent to $q_i^{\emptyset}$ at the same rate as it is sent to q_i, for all i 1;

(ii) *allocation of objects to agents within a queue is according to a scaled version of* s*:* there is $p_i > 0$ such that, for all $x \in [0, p_i]$, the probability that an agent is matched before reaching position x of $q_i^{\emptyset}$ equals $1 - \mathsf{s}_{p_i}(x)$. The agent's position in the queue as a function of her waiting time is given by a time index p_i with $\mathsf{p}_i(0) = p_i$. (That is, every agent who joins $q_i^{\emptyset}$ starts at position p_i.) Moreover, conditional on being matched at a position, the object type allocated to the agent has the same distribution as the distribution of the object types sent to $q_i^{\emptyset}$; and

(iii) *an agent is sent to a queue that maximizes her payo :* The payoff of an agent from joining a queue is well-defined, according to (I.6). Every agent is sent to a queue that maximizes her payoff.

The above conditions mirror those in the definition of s-alteration in the previous subsection. Condition (i) above is imposed there identically, and condition (ii) is counterpart to the detail-freeness condition imposed there. Condition (iii) guarantees incentive compatibility.

Example I.1 (SIRO-alterations of a mechanism)**.** *Let* s_U *be the uniform distribution over the unit interval, and* $N^{\emptyset}$ *be an* s_U*-alteration of an indirect queue-based mechanism* N*. Then, every object sent to a queue in* $N^{\emptyset}$ *is allocated uniformly at random to one of the agents present in that queue.*

Definition I.4. *A normalized queuing scheme* s *is* universally optimal *if for every market* M *and an optimal queue-based mechanism* N *for that market, there exists an* s*-alteration of* N *that is an optimal mechanism for market* M*.*

Proposition I.3. *No normalized queuing scheme is universally optimal.*

Proof. The proof is by contradiction. Let s be a normalized queuing scheme that is universally optimal. For $\beta \in \mathsf{R}_+$, let p_β be the unique time index consistent with s_β such that $\mathsf{p}_\beta(0) = \beta$. The existence and uniqueness of p_β is due to Proposition I.1. Also, let t_β be the inverse function of p_β. The inverse function exists as p_β is continuous and strictly increasing, by the same proposition. We denote $\mathsf{s}_1, \mathsf{p}_1, \mathsf{t}_1$ by $\mathsf{s}, \mathsf{p}, \mathsf{t}$.

Choose $p_1 \in (0,1)$ with $\mathsf{s}(p_1) \in (0,1)$. Let $\delta_1 = \mathsf{s}(p_1)$, and $p_2 \in (0, p_1)$ be such that $\mathsf{s}(p_1) - \mathsf{s}(p_2) < \delta_1(1 - \delta_1)$. The existence of such p_2 follows from the right-continuity of the CDF $\mathsf{s}(\,)$. Let

$$
\begin{aligned}
&t_1 = \mathsf{t}(p_1),\ t_2 = \mathsf{t}(p_2),\\
&\epsilon = \mathsf{s}(p_1) - \mathsf{s}(p_2),\ \delta_2 = 1 - \delta_1 - \epsilon,\\
&L = 1,\ H = \frac{1+\delta_1}{1 - \frac{\epsilon}{1-\delta_1} - \delta_2 - \epsilon} + 1.
\end{aligned}
$$

Observe that $H > L$, because $\epsilon < \delta_1(1 - \delta_1)$. Choose c to be a cost function satisfying

$$
\begin{aligned}
&c(t) = L \text{ for } t < t_1 \text{ and } c(t_1) = L, \text{ and}\\
&c(t) = H \text{ for } t > t_2 \text{ and } c(t_2) = H.
\end{aligned}
$$

For any $\alpha > 0$, define the cost function c_α by $c_\alpha(t) = \alpha c(t)$ for all $t \in \mathbb{R}_+$.

Let M be an arbitrary market where the waiting cost function is c. Also, let M be an optimal monotone disjoint queue mechanism for M. Denote the queues in M by $q_0, q_1, \ldots, q_k$, and let $\bar{c}_i$ be the expected waiting cost at q_i. We choose M such that $\bar{c}_i > 0$ and a positive rate of agents join q_i for $i \geq 1$. Let N be an s-alteration of M that is optimal for M. Denote the queues in N by $q_0^0, q_1^0, \ldots, q_k^0$. Let $\bar{p} > 0$ and $(\mathsf{s}_{\bar{p}}, \bar{p})$ be the queuing schedule used in q_1^0. (Thus, $\bar{p}$ is the position of an agent immediately after joining q_1^0.)

For $\alpha \in (0, 1]$, recall that $\mathsf{t}_{\bar{p}}(\alpha\bar{p})$ gives the time that it takes for an agent who joins q_1^0 to reach position $\alpha\bar{p}$ from position $\bar{p}$ (conditional on not being matched before reaching position $\alpha\bar{p}$). By Proposition I.1,

$$\mathsf{t}_{\bar{p}}(\alpha\bar{p}) = \int_{\alpha\bar{p}}^{\bar{p}} \frac{1}{\mathsf{s}_{\bar{p}}(x)}\,\mathrm{d}x = \int_{\alpha\bar{p}}^{\bar{p}} \frac{1}{\mathsf{s}(x/\bar{p})}\,\mathrm{d}x = \int_{\alpha}^{1} \frac{1}{\mathsf{s}(z)}\bar{p}\,\mathrm{d}z,$$

where the last equality follows from a change of variables $z = x/\bar{p}$. Let t denote t_1. By the above equation, $\mathsf{t}_{\bar{p}}(\alpha\bar{p}) = \bar{p}\,\mathsf{t}(\alpha)$. Then, we can write the expected waiting cost of joining q_1^0 as

$$\int_0^{\bar{p}} c(\mathsf{t}_{\bar{p}}(p))\,\mathrm{d}\mathsf{s}_{\bar{p}}(p)) = \int_0^1 c(\bar{p}\,\mathsf{t}(p/\bar{p}))\,\mathrm{d}\mathsf{s}_{\bar{p}}(\alpha\bar{p}) = \int_0^1 c(\bar{p}\,\mathsf{t}(\alpha))\,\mathrm{d}\mathsf{s}(\alpha). \tag{I.7}$$

The first equality follows from a change of variables $p = \alpha\bar{p}$. By the envelope theorem, this also equals the expected waiting cost from joining q_1. We next observe that there is $\kappa > 0$ such that

$$\int_0^1 c(\bar{p}\,\mathsf{t}(\alpha))\,\mathrm{d}\mathsf{s}(\alpha) = \int_0^1 \kappa c(\mathsf{t}(\alpha))\,\mathrm{d}\mathsf{s}(\alpha) = \int_0^1 c_\kappa(\mathsf{t}(\alpha))\,\mathrm{d}\mathsf{s}(\alpha). \tag{I.8}$$

The first equality holds since the middle expression is an increasing function of κ, and approaches 0 (infinity) as κ approaches 0 (infinity). The second equality holds by the definition of c_κ.

Let M_κ be a market that is identical to M but for its cost function being c_κ (instead of c in M). By Proposition I.2, a monotone disjoint queue mechanism exists for M_κ such that (i) the mechanism has a queue q_i^{00} for each queue q_i in M, (ii) each object type is sent to q_i^{00} at the same rate as it is sent to q_i, for all i, and (iii) M_κ is an optimal mechanism for M_κ.

Since s is a universally optimal normalized queuing scheme, then M_κ has an s-alteration which is optimal for M_κ. We denote this optimal mechanism by N_κ. A necessary condition for the optimality of N_κ is incentive compatibility. From the envelope condition it follows that the expected waiting cost in q_1^{00} equals the expected waiting cost in q_1. Hence, the expected waiting cost in q_1^{00} is $\int_0^1 c_\kappa(\mathsf{t}(\alpha))\,\mathrm{d}\mathsf{s}(\alpha)$, by (I.7) and (I.8). Consequently, the queuing scheme used in q_1^{00} is $(\mathsf{s}, 1)$ and $\mathsf{t}(\alpha)$ gives the time that agents at position α have waited thus far. We next show that the continuation payoff of the agents at position p_1 would be negative, which would contradict individual rationality, and thus optimality of N_κ.

Let t be a random variable denoting the time that an agent who has joined q_1^{00} receives an object. The expected waiting cost of joining q_1^{00} is

$$\delta_1 \mathsf{E}_t\ [c_\kappa(t\ )|t \quad t_1] + \epsilon\, \mathsf{E}_t\ [c_\kappa(t\ )|t\ \ 2\ (t_1, t_2]] + \delta_2\, \mathsf{E}_t\ [c_\kappa(t\ )|t\ \ > t_2]\,. \tag{I.9}$$

On the other hand, the waiting cost that an agent at position p_1 expects to incur, excluding the cost she has incurred thus far, is at least

$$\frac{\epsilon}{1 \quad \delta_1} \mathsf{E}_t\ [c_\kappa(t\ ) \quad c_\kappa(t_1)|t\ \ 2\ (t_1, t_2]] + \frac{1 \quad \epsilon \quad \delta_1}{1 \quad \delta_1} \mathsf{E}_t\ [c_\kappa(t\ ) \quad c(t_1)|t\ \ > t_2]\,. \tag{I.10}$$

We will show that (I.10) is strictly larger than (I.9). That would mean that an agent of type θ_0 who attains an expected payoff of 0 from joining q_1^{00} (i.e., is indifferent between joining q_0^{00} and q_1^{00} upon arrival) has a negative continuation payoff at position p_1, which contradicts individual rationality. In fact, as the utility function u is continuous in θ, this argument implies that there is $\theta_1 > \theta_0$ such that the latter statement holds for all agents of type belonging to $[\theta_0, \theta_1)$. It remains to show that (I.10) is strictly larger than (I.9) .

To this end, let $\bar{H} = \mathsf{E}_t\ [c(t\ )|t\ \ > t_2]$. Note that $\bar{H} \quad H = c(t_2)$, and observe that

$$\frac{\bar{H}}{L} > \frac{1 + \delta_1}{1 \quad \frac{\epsilon}{1 \quad \delta_1} \quad \delta_2 \quad \epsilon}$$

$$) (1 \quad \frac{\epsilon}{1 \quad \delta_1} \quad \delta_2)\bar{H} > \epsilon\bar{H} + (1 + \delta_1)L$$

$$) (1 \quad \frac{\epsilon}{1 \quad \delta_1})(\bar{H} \quad L) > \delta_1 L + \epsilon H + \delta_2 \bar{H}$$

$$) \frac{1 \quad \epsilon \quad \delta_1}{1 \quad \delta_1} \mathsf{E}_t\ [c_\kappa(t\ )|t\ \ > t_2] > \delta_1 \mathsf{E}_t\ [c_\kappa(t\ )|t \quad t_1] + \epsilon\, \mathsf{E}_t\ [c_\kappa(t\ )|t\ \ 2\ (t_1, t_2]] + \delta_2\, \mathsf{E}_t\ [c_\kappa(t\ )|t\ \ > t_2]\,.$$

Hence, (I.9) is strictly smaller than (I.10), which is the promised claim. □

II Proofs from Section 5

II.1 Preliminaries

Recall from (3.1) that the planner's objective can be written as

$$\max_{X\, 2 \mathrm{MPS}(X_{\mathrm{PAM}})} \int_0^{\bar{\theta}} X(\theta)\, \underbrace{(\lambda_E \theta + \lambda_W \phi_G(\theta))}_{\psi(\theta)}\, \mathrm{d}G(\theta). \tag{II.1}$$

This problem can be solved using classical ironing techniques. In particular, we use Proposition 2 of Kleiner et al. (2021), as follows. Define $\Psi : [0,1]\ \ !\ \ \mathsf{R}$ such that $\Psi(q) = \int_0^q \psi(G\ \ ^1(q))\, \mathrm{d}q$. Let $\overline{\Psi}$ denote the convex envelope of Ψ, i.e., the largest convex function that lies below Ψ.

Proposition 0 (Proposition 2 of Kleiner et al. (2021)). *X is an optimal solution to (II.1) if there exists a family I of disjoint intervals $\{[\alpha_i, \beta_i)\}_{i \in I}$ with $0 \leq \alpha_i < \beta_i \leq 1$ for all i such that*

(i) $\overline{\Psi}$ is a concave over $[\alpha_i, \beta_i)$ for all $i \in I$ and $\overline{\Psi}(q) = \Psi(q)$ if q belongs to no interval in I, and

(ii) for $\theta \in \Theta$, $X(\theta) = \frac{\int_{\alpha_i}^{\beta_i} X_{\text{PAM}}(G^{-1}(q))dq}{\beta_i - \alpha_i}$ if $G(\theta) \in [\alpha_i, \beta_i)$ for some $i \in I$; otherwise, $X(\theta) = X_{\text{PAM}}(\theta)$.

II.2 Proof of Proposition 5.1

Let X_N denote the positive assortative assignment X_{PAM} when the objects' arrival rate is N. First, we compute the derivative

$$\psi'(\theta) = \lambda_E + \lambda_W \left(\frac{(G(\theta) - 1)g'(\theta)}{g(\theta)^2} \quad 1 \right), \tag{II.2}$$

which exists for $\theta \in \Theta$ since g is continuously differentiable and has full support over $[0, \bar{\theta}]$. Hence,

$$\lim_{\theta \uparrow \bar{\theta}} \psi'(\theta) = \lambda_E - \lambda_W. \tag{II.3}$$

This holds because g is continuously differentiable over its support and thus has bounded derivatives, and $\lim_{\theta \uparrow \bar{\theta}} \frac{(G(\theta) - 1)g'(\theta)}{g(\theta)^2} = 0$. By the two latter equations, for any $\epsilon > 0$ there is $\theta_\epsilon < \bar{\theta}$ such that if $\lambda_W < \lambda_E - \epsilon$ ($\lambda_E < \lambda_W - \epsilon$) then ψ is increasing (decreasing) over $[\theta_\epsilon, \bar{\theta}]$. It follows that Ψ is convex (concave) over $[\theta_\epsilon, \bar{\theta}]$ if $\lambda_W < \lambda_E - \epsilon$ ($\lambda_E < \lambda_W - \epsilon$). We next consider two cases.

If $\lambda_W < \lambda_E - \epsilon$, then Ψ is convex over $[G(\theta_\epsilon), 1]$. Hence, $\Psi(q) = \overline{\Psi}(q)$ for $q > G(\theta_\epsilon)$. Moreover, there is a sufficiently small $N_\epsilon > 0$ such that $X_N(\theta_\epsilon) = \omega_0$ when $N < N_\epsilon$. The two latter facts, together with Proposition 0, imply that $X = X_N$ is a solution to (II.1) when $N < N_\epsilon$. Thus, a positive assortative assignment is optimal, which can be implemented by a full-disclosure policy.

On the other hand, if $\lambda_E < \lambda_W - \epsilon$, then Ψ is strictly concave over $[G(\theta_\epsilon), 1]$. Let θ_ϵ^0 be such that $(G(\theta_\epsilon^0), 1)$ is the maximal open interval over which Ψ is strictly concave. Thus, we have $\theta_\epsilon^0 \geq \theta_\epsilon$ and $\Psi(q) \leq \overline{\Psi}(q)$ for $q \in (G(\theta_\epsilon^0), 1)$. When $N < N_\epsilon$, then $X_N(\theta_\epsilon) = \omega_0$. From the two latter facts and Proposition 0 it follows that

$$X(\theta) = \begin{cases} \omega_0, & \text{if } \theta > \theta_\epsilon^0, \\ \frac{\int_{G(\theta_\epsilon^0)}^1 X_N(G^{-1}(q))\mathrm{d}q}{1 - G(\theta_\epsilon^0)}, & \text{otherwise} \end{cases} \tag{II.4}$$

is an optimal solution to (II.1). This can be implemented by a disclosure policy that has a realization space $\{s_0, s_1\}$, where receiving s_0 means the object has type ω_0 with probability 1, and s_1 is disclosed for every object type $\omega \in \Omega_+$. (Notably, it is possible that some of the objects of type ω_0 are pooled with the objects of type Ω_+). In that case s_1 is also sent with positive probability for objects of type ω_0.) This is a no-disclosure policy, since the realization disclosed for every object type in Ω_+

is the same.

II.3 Proof of Proposition 5.2

A function $z : [0,1] \; ! \; \mathsf{R}$ is *convex-concave* if there is $\tilde{q} \quad 0$ such that $z(q)$ is convex for $q \quad \tilde{q}$ and concave for $q \quad \tilde{q}$. Similarly, z is *concave-convex* if there is $\tilde{q}$ such that z is concave for $q \quad \tilde{q}$ and convex over $q \quad \tilde{q}$.

When the planner's objective is welfare-maximization, $\lambda_E = 0$ and $\psi(\theta) = \frac{1}{h_G(\theta)}$. From the fact that the function $\frac{1}{h_G}$ is single-peaked it follows that Ψ is convex-concave. Recall that X_N denotes the positive assortative assignment when the objects' arrival rate equals N. When Ψ is convex-concave, then by Proposition 0 there is $\theta \; 2 \; \Theta$ such that

$$X_N(\theta) = \begin{cases} X_N(\theta), & \theta < \theta \; , \\ \frac{\int_{G(\theta \;)}^{1} X_N(G \;\; ^{1}(q))\mathrm{d}q}{1 \;\; G(\theta \;)} & \theta \quad \theta \end{cases} \tag{II.5}$$

is an optimal solution to (II.1). That is, the optimal solution pools agent types above θ , but does not pool other agent types. This can be seen from Figure 8. Note that X_N is an extreme point of MPS(X_N), by Lemma 0. The policy $\mu(N)$ is constructed so that it pools together the object types above $X_N(\theta \;)$ and fully discloses the object types below $X_N(\theta \;)$. For an object of type $X_N(\theta \;)$, the policy fully discloses the type with a fixed probability $p_N > 0$, and with probability $1 \quad p_N$ does not disclose the object type (i.e., pools it together with objects of type above $X_N(\theta \;)$). By construction, $\mu(N)$ is an upper-censorship policy. (Figure 10 demonstrates such a disclosure policy and its corresponding monotone disjoint queue mechanism.) Since $X_N(\theta \;)$ decreases as N goes down, then $\mu(N)$ is less informative for lower values of N.

It remains to consider the case where $\frac{1}{h_G}$ is single-dipped. Then, Ψ is concave-convex. The proof in this case is similar, with the difference that the optimal solution X_N is given by

$$X_N(\theta) = \begin{cases} \frac{\int_{0}^{G(\theta \;)} X_N(G \;\; ^{1}(q))\mathrm{d}q}{G(\theta \;)}, & \theta < \theta \; , \\ X_N(\theta) & \theta \quad \theta \; . \end{cases}$$

The disclosure policy $\mu(N)$ pools together the object types below $X_N(\theta \;)$ and fully discloses the object types above $X_N(\theta \;)$. Thus, it is a lower-censorship policy. Since $X_N(\theta \;)$ decreases as N goes down, then the policy is more informative for lower values of N.

III Proofs and simulations for Section 6

Throughout the proof, we refer to the setup of Section 2 as the *continuum model*, and the discrete setup of Section 6 as the *discrete model*. Let H_s denote the history of the process in the discrete

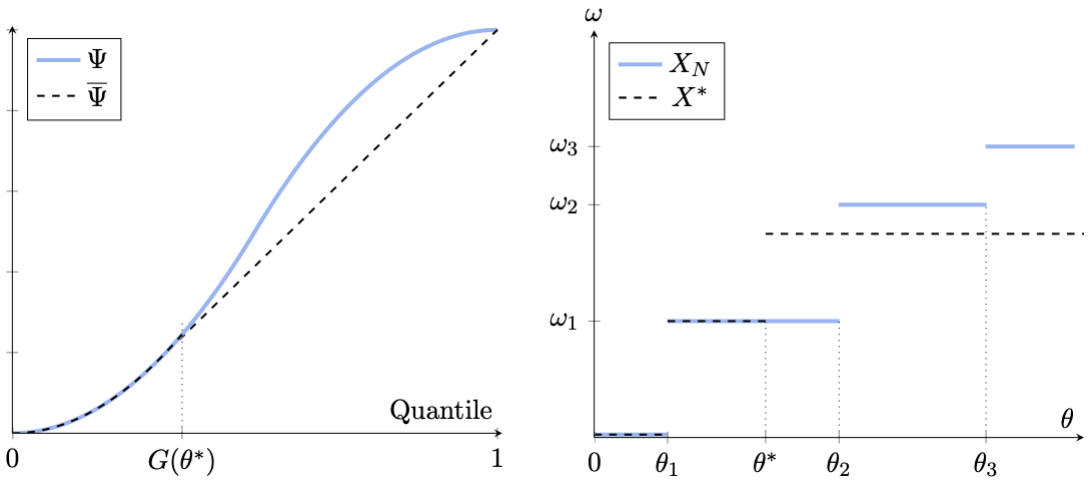


Figure 8: **Left panel:** Ψ is convex-concave. **Right panel:** $X(\theta)$ equals $\frac{\int_{G(\theta)}^{1} X_N(G^{-1}(q))\mathrm{d}q}{1 - G(\theta)}$ for $\theta > \theta$ and equals $X_N(\theta)$ otherwise. The corresponding monotone disjoint queue mechanism sends objects of type ω_0 to q_0, objects of type ω_1 to q_1 at a rate $G(\theta) - G(\theta_1)$, and pools the remaining fraction of objects of type w_1 with all objects of type w_2 and ω_3 and sends them together to q_2.

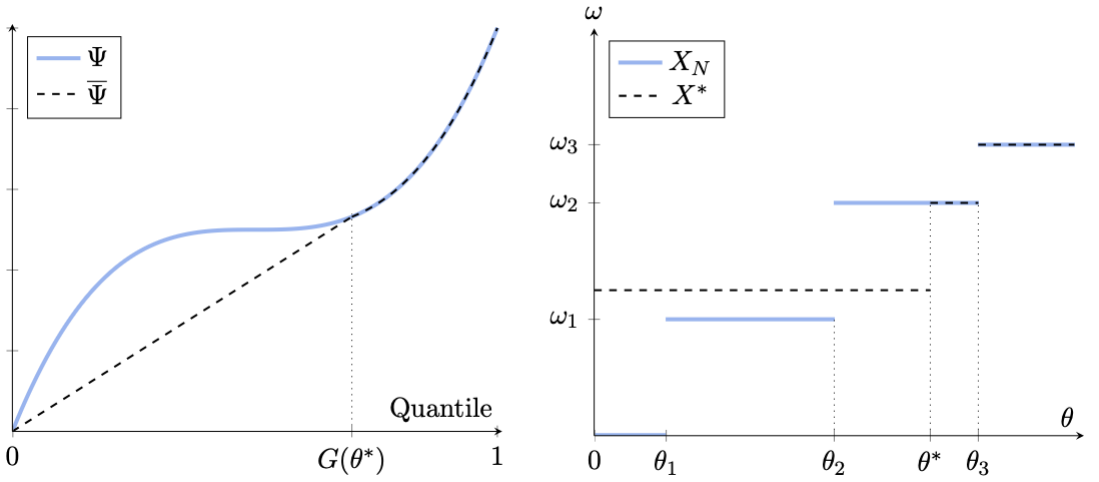


Figure 9: **Left panel:** Ψ is concave-convex. **Right panel:** the optimal solution $X(\theta)$. The corresponding monotone disjoint queue mechanism sends to q_1 objects of type $\omega_0, \omega_1, \omega_2$ respectively at rates $G(\theta_1), F(\omega_2), G(\theta) - G(\theta_2)$ objects of type ω_2 to q_2 at rate $G(\theta_3) - G(\theta)$, and objects of type ω_3 to q_3 at rate $F(\omega_3)$.

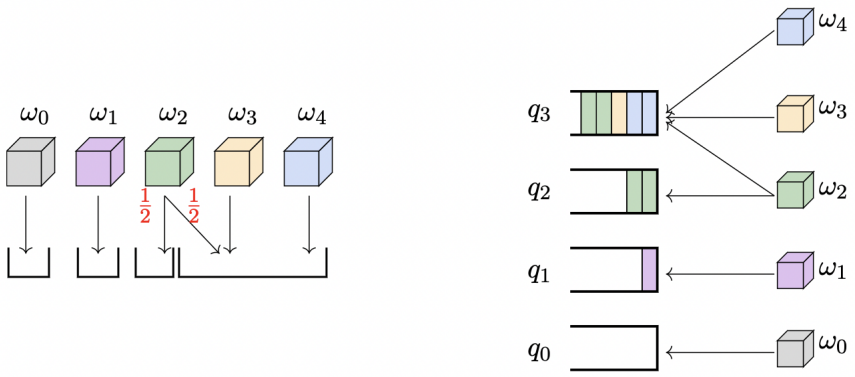


Figure 10: Example of an upper-censorship disclosure policy (**left**) and its implementation via a monotone disjoint queue mechanism (**right**). A fraction of objects with type ω_2 are pooled together of types ω_3 and ω_4.

model from time 0 until time s 0. For a mechanism M in the discrete model let the random variable $A_M^{H(s)}(\theta)$ denote the allocation timeline that the mechanism promises to an agent who arrives at time s and has type θ. (We note that $A_M^{H(s)}(\theta)$ can depend on the history $H(s)$, and thus is a random variable.) Let $X_M^{H(s)}(\theta)$ denote the expected object type that is assigned to the agent according to the allocation timeline $A_M^{H(s)}(\theta)$. For every s, we let $X_M^s(\theta)$ be the random variable that takes the value $X_M^{H(s)}(\theta)$ when the history at time s is $H(s)$. Define $C_M^{H(s)}(\theta)$ and $C_M^s(\theta)$ similarly, but for the expected waiting cost of the agent (instead of the expected object type assigned to her).

Definition III.1. *A mechanism M is a* steady-state *mechanism if the following two conditions hold: First, for every θ 2 Θ there exist CDFs $D_M(\theta), Y_M(\theta)$ such that $C_M^s(\theta)$ and $X_M^s(\theta)$ converge in distribution respectively to $D_M(\theta)$ and $Y_M(\theta)$, as s approaches in nity. Denote the means of $D_M(\theta), Y_M(\theta)$ by $c_M(\theta), X_M(\theta)$, respectively. Second, $c_M(\,), X_M(\,)$. are Lebesgue integrable.*

Throughout we refer to c_M, X_M respectively as the interim payment and interim allocation rules of M. Also, we sometimes say that M has a steady state when it is a steady-state mechanism.

Lemma III.1. *For a steady-state mechanism M, X_M 2 MPS$_{\mathrm{w}}(X_{\mathrm{PAM}})$.*

Proof. The fact that X_M is increasing follows from the fact that ex post incentive compatibility implies interim incentive compatibility. The fact that X_{PAM} weakly majorizes X_M follows from the same argument as in the proof of Lemma A.4. □

Proof of Proposition 6.1. Throughout the proof, by a *mechanism* we mean a steady-state mechanism. For every X 2 MPS(X_{PAM}), define $V(X) = \lambda_E E(X) + \lambda_W W(X)$. Also, for any market size ζ, let the *second-best welfare*, denoted by $W(\zeta)$, be the supremum over the welfares of all incentive-compatible and individually rational mechanisms. Also, let $\tilde{W}(\zeta)$ be the supremum over

the welfares of all monotone disjoint queue mechanisms. Define $E(\zeta)$ and $\tilde{E}(\zeta)$ similarly but for the notion of efficiency (instead of welfare).

Let $X^* = \arg\max_{X \in \mathrm{MPS}_{\mathrm{w}}(X_{\mathrm{PAM}})} V(X)$. In the proof of Theorem 3.1 we showed that X^* can be chosen so that it is an extreme point of $\mathrm{MPS}(X_{\mathrm{PAM}})$. Moreover, we showed that such extreme point X^* corresponds to a monotone disjoint queue mechanism in the continuum model. Denote this mechanism by M^*. For every market size ζ, $V(\zeta) \leq V(X^*)$, since the interim allocation rule of every mechanism in the discrete model belongs to $\mathrm{MPS}_{\mathrm{w}}(X_{\mathrm{PAM}})$, by Lemma III.1. We next construct a mechanism $\tilde{M}(\zeta)$ for every $\zeta > 0$ such that the planner's objective under $\tilde{M}(\zeta)$ approaches $V(X^*)$ as ζ approaches infinity.

Suppose that there are k disjoint queues in M^*, and let $f_{\omega,i}$ denote the fraction of objects of type ω that are sent to queue i. In the discrete model, for any market size ζ, let $\tilde{M}(\zeta)$ denote the disjoint queue mechanism in which every object of type ω, upon its arrival, is sent to one of the queues $1, \ldots, k$ independently, such that the object is sent to queue i with probability $f_{\omega,i}$. Proposition 3 of Ashlagi et al. (2022a) provides concentration bounds for the discrete model. From this proposition it directly follows that $\tilde{M}(\zeta)$ has a steady state, and as ζ approaches infinity,

(i) the steady-state (i.e., time-average) waiting cost from joining queue i under $\tilde{M}(\zeta)$ approaches the waiting cost at queue i under M^*, and

(ii) for all but a measure zero set of agent types $\theta \in \Theta$,[6] the average object type assigned to an agent of type θ under $\tilde{M}(\zeta)$ approaches the average object type assigned to an agent of type θ under M^*.

By (i) and (ii), for all but a measure zero set of agent types $\theta \in \Theta$, the steady-state payoff of an agent of type θ under $\tilde{M}(\zeta)$ approaches the steady state payoff of an agent of type θ under M^*, as ζ approaches infinity. Therefore, $\lim_{\zeta \to \infty} \tilde{W}(\zeta) = W(X^*)$. This fact, together with $\tilde{W}(\zeta) \leq W(\zeta) \leq W(X^*)$, implies that $\lim_{\zeta \to \infty} \tilde{W}(\zeta) = \lim_{\zeta \to \infty} W(\zeta)$. By a similar argument, (i) implies that $\lim_{\zeta \to \infty} \tilde{E}(\zeta) = \lim_{\zeta \to \infty} E(\zeta)$. The claim of the theorem follows the two latter equalities. □

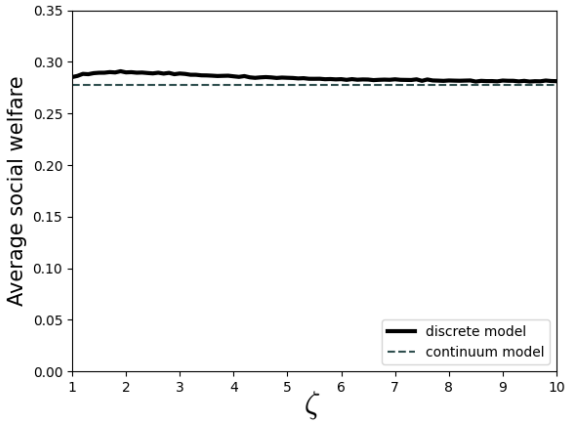
III.1 Simulations for moderately small markets

We next run simulations to investigate the convergence rate of the above limit results. We consider a disjoint queue mechanism with two queues in the discrete setup. These simulations indicate convergence for moderately small values of ζ. Let G be the uniform distribution over the unit interval. Suppose that objects arriving at queue i have an expected type of $\bar{\omega}_i = i$ and an arrival rate of ζ, for $i \in \{1, 2\}$. Agents arrive according to a Poisson process at rate 3ζ.

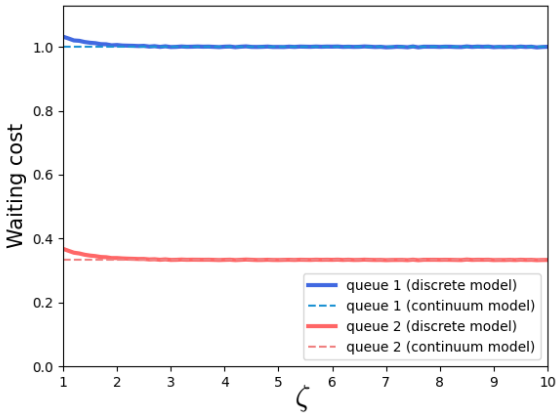
We set $\gamma = 0.05$ and vary ζ from 1 to 10 with increments of 0.1. For every such ζ, we run the process until 10^7 agents arrive to the market, and then compute the average payoff of these agents. The result is plotted in Figure 11a. In addition, Figure 11b plots the expected cost that

[6] The measure zero set is the set of agent types that are *indifferent* between joining at least two of the queues under the optimal monotone disjoint queue mechanism in the continuum model.

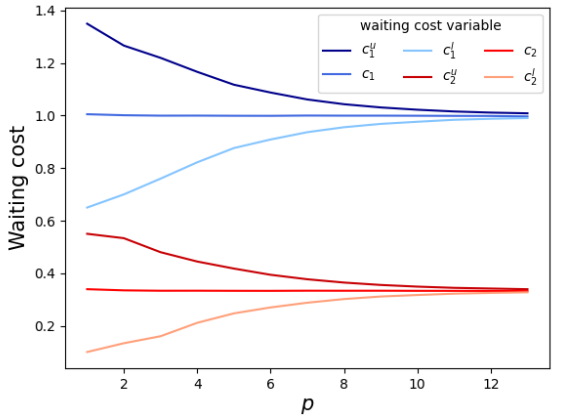
an agent incurs if she joins each of the queues upon her arrival. We observe that the social welfare and the expected waiting cost at each queue approach their counterpart values in the continuum model where agents arrive with a flow rate of 3ζ and objects with expected type $\bar{\omega}_i$ arrive at queue i with a flow rate of ζ, for i *2 f*$1, 2g$.



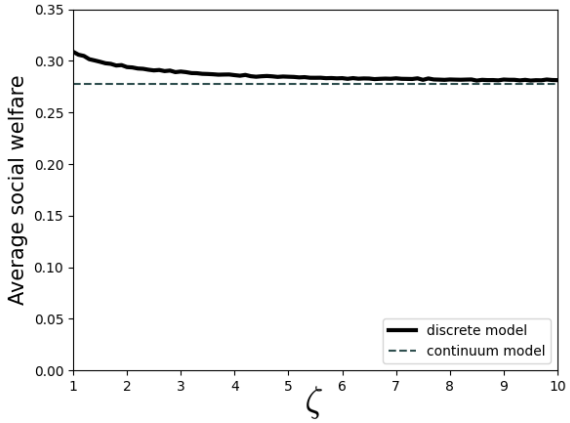
(a) Average social welfare in ζ.



(b) Average expected waiting cost in ζ.



(c) Concentration bounds: $\zeta = 10 \quad 2^p$.



(d) Average social welfare: objects are queued.

Figure 11

Figure 11c shows that the expected waiting costs that an agent faces from joining each queue upon her arrival are sharply concentrated around their means. This is demonstrated in the figure by plotting, for every arrival rate ζ, an interval $[c_i^l, c_i^u]$ for each queue i. The interval is chosen so that, for 95% of the agents, it contains the waiting cost that the agent expects to incur if she joins queue i (i.e., the interval excludes 2.5% of the data points from each tail of the distribution). The intervals shrink as λ grows large; i.e., waiting costs become concentrated around the means.

Finally, we consider a slightly different discrete setting where, if an object arrives at a queue

when that queue is empty, the object is not discarded. Instead, we assume that such objects are preserved for the agents who choose to join that queue in the future, and will be assigned to these agents in a first-come, first-served manner. We make similar observations in this setting. In particular, in Figure 11d we observe that social welfare in the discrete model approaches social welfare under the continuum model as ζ grows large (an observation analogous to that shown in Figure 11a).

IV Examples

IV.1 Pareto frontier derivation for Section 2.1

In this section, we derive the Pareto frontier for the example considered in Section 2.1. Recall from (II.1) that the planner's objective is

$$\int_0^1 \left(\lambda_E \theta X(\theta) + \lambda_W \frac{1 \quad G(\theta)}{g(\theta)} X(\theta)\right) \mathrm{d}G(\theta) = \int_0^1 ((\lambda_E \quad \lambda_W)\theta + \lambda_W) X(\theta)\, \mathrm{d}\theta. \tag{IV.1}$$

Setting $\psi(\theta) = (\lambda_E \quad \lambda_W)\theta + \lambda_W$ and $\Psi(q) = (\lambda_E \quad \lambda_W)\frac{\theta^2}{2} + \lambda_W \theta$, we can apply Proposition 0 to solve the reduced form of the planner's problem, as follows.

If $\lambda_E = \lambda_W$, then ψ is a nonnegative constant function. It follows that any X *2* MPS(X_{PAM}) attains the same objective for the planner. Thus, the Pareto frontier is a 45-degree line segment, as in Figure 2. If $\lambda_E > \lambda_W$ then $\Psi(q)$ is convex therefore $\Psi(q) = \overline{\Psi}(q)$. Thus, the planner's objective (IV.1) is maximized over MPS(X_{PAM}) at $X = X_{\text{PAM}}$. Note that $X = X_{\text{PAM}}$ corresponds to the choice of $\epsilon = 0$. On the other hand, if $\lambda_E < \lambda_W$, then $\Psi(q)$ is concave. By Proposition 0, the solution that for all θ sets $X\ (\theta) = \int_0^1 X_{\text{PAM}}(\theta) d\theta = (\omega_1 + \omega_2)/2$ attains the highest objective for the planner. This solution corresponds to complete pooling, i.e., $\epsilon = \frac{1}{2}$. Therefore, the two endpoints of the 45-degree line segment correspond to the cases of $\epsilon = 0$ and $\epsilon = \frac{1}{2}$, respectively.

IV.2 Example: Dynamic IC and IR do not hold under SIRO

Consider an arbitrary market and let M_{SIRO} be a mechanism in that market that is a monotone disjoint queue mechanism with the difference that, within each queue, the objects are allocated to agents according to a SIRO rule: an arriving object is allocated to an agent in the queue who is drawn uniformly at random. Below we show that dynamic IC fails under SIRO for natural examples of convex and concave cost functions. We also show that (dynamic) IR fails under SIRO when the cost function is strongly convex.

Fact IV.1. *M_{SIRO} is not individually rational if c is strongly convex. If, in addition, M_{SIRO} involves at least two queues, then it is not dynamically incentive-compatible.*

Proof. Let t_i be the random variable that gives the time of allocation for an agent after joining queue i. Since SIRO is used in each queue, then t_i, t_j are distributed exponentially (but with different means). Consider an agent of type θ who has joined queue i and has waited t units of time. Then, for M_{SIRO} to be dynamically incentive-compatible, for every queue j and $t > 0$, we must have

$$u(\theta, \overline{\omega}_i) \quad \mathsf{E}_{t_i}\left[c(t_i + t) \quad c(t)\right] \quad u(\theta, \overline{\omega}_j) \quad \mathsf{E}_{t_j}\left[c(t_j + t) \quad c(t)\right], \tag{IV.2}$$

where $\overline{\omega}_i, \overline{\omega}_j$ are the expected object types sent to queues i, j, respectively. Choose i, j such that $\overline{\omega}_i > \overline{\omega}_j$. Hence, the mean of t_i is larger than the mean of t_j. Since t_i, t_j are exponentially distributed, then t_i first-order stochastically dominates t_j. When c is strictly convex, it follows that $\mathsf{E}_{t_i}\left[c(t_i + t) \quad c(t)\right] \quad \mathsf{E}_{t_j}\left[c(t_j + t) \quad c(t)\right]$ approaches infinity as t does. Thus, (IV.2) fails to hold for a positive measure of agents who join queue i. That is, the mechanism is not dynamically incentive-compatible.

To prove the claim about individual rationality, consider the lowest type queue, queue 1, and an agent of type θ who has joined that queue and has not been matched after t units of time. For M_{SIRO} to be individually rational, it must hold for every $t > 0$ that

$$u(\theta, \overline{\omega}_1) \quad \mathsf{E}_{t_1}\left[c(t_1 + t) \quad c(t)\right] \quad 0.$$

When c is strongly convex, $\mathsf{E}_{t_1}\left[c(t_1 + t) \quad c(t)\right]$ approaches infinity with t. Therefore, the above equation fails to hold for a positive measure of agents. Thus, M_{SIRO} is not individually rational. □

Remark IV.1. *Failure of dynamic incentive-compatibility does not hinge on convexity of the cost function. For example, suppose that* $c(t) = \sqrt{p}\,\bar{t}$. *Then, in the above proof we can choose* i, j *such that* $\overline{\omega}_j > \overline{\omega}_i$. *Then,* $\mathsf{E}_{t_i}\left[c(t_i + t) \quad c(t)\right] \quad \mathsf{E}_{t_j}\left[c(t_j + t) \quad c(t)\right]$ *approaches zero as* t *approaches in nity, because* t_j *rst-order stochastically dominates* t_i. *Hence, agents who remain in queue* i *for a su ciently long time prefer to switch to* q_j *instead. Therefore, the mechanism is not dynamically incentive-compatible.*

IV.3 Example: FCFS policies with full or no information can be suboptimal

Example IV.1. Suppose that $\Theta = [0, \bar{\theta}]$, $\Omega = \{\omega_1, \omega_2\}$, $F(\omega_1) = F(\omega_2) = 0.25$, $u(\theta, \omega) = \theta\omega$, and $N = 0.5$. Denote the upper quartile of distribution G by q and the median of G by m.

Let M denote the mechanism that uses no pooling and $M^{\emptyset}$ be the mechanism that uses complete pooling. We next compute the social welfare in M and $M^{\emptyset}$, respectively. Under M, all agents that are assigned to an object of the same type wait in the same queue and thus have the same waiting cost. Let c_1, c_2 respectively denote the waiting costs for objects of type ω_1, ω_2. An agent of type q must be indifferent between accepting objects of type ω_1 and objects of type ω_2, whereas an agent

of type m must be indifferent between accepting objects of type ω_1 and outside option ω_0. Thus,

$$u(q,\omega_2) \quad c_2 = u(q,\omega_1) \quad c_1,\ u(m,\omega_1) \quad c_1 = 0.$$

It follows that $t_1 = m\omega_1$ and $t_2 = q(\omega_2 \quad \omega_1)+m\omega_1$. Consequently, every agent of type $\theta > q$ accepts only objects of type ω_2, and therefore has a payoff $V_{M}(\theta) = \theta\omega_2 \quad q(\omega_2 \quad \omega_1) \quad m\omega_1$. On the other hand, agents of type θ 2 (m,q) accept only objects of type ω_1 and get payoff $V_{M}(\theta) = \theta\omega_1 \quad m\omega_1$. Since an agent of type q 2 Θ is indifferent between being assigned to an object ω_1 and the outside option, then $t_0 = 0$. Hence, every agent of type $\theta < m$ has a payoff $V_{M}(\theta) = 0$. Therefore, the social welfare under M is

$$W_{M} = \frac{1}{4}\left(\mathsf{E}_{\theta}\left[\theta \,j\, \theta > q\right]\omega_2 \quad q(\omega_2 \quad \omega_1) \quad m\omega_1\right) + \frac{1}{4}\left(\mathsf{E}_{\theta}\left[\theta \,j\, m < \theta < q\right]\omega_1 \quad m\omega_1\right).$$

To compute the social welfare under M^{0}, we observe that in M^{0} the payoff of agents of type θ is $u(\theta,\omega_2)F(\omega_2) + u(\theta,\omega_1)F(\omega_1) = \theta(\omega_2+\omega_1)/4$. The social welfare in M^{0} then is

$$W_{M^{0}} = \mathsf{E}_{\theta}\left[\theta\right](\omega_1+\omega_2)/4.$$

We observe that when, e.g., G is the arcsine distribution over the unit interval,

$$W_{M} = 0.136 \quad \omega_1 + 0.024 \quad \omega_2 < 0.125 \quad \omega_1 + 0.125 \quad \omega_2 = W_{M^{0}},$$

Therefore, the complete-pooling mechanism M^{0} achieves higher social welfare than the no-pooling mechanism M.

We also show that there is another mechanism M^{00} that achieves higher welfare than both M, M^{0}. Consider the mechanism M^{00} that randomly allocates all objects of type ω_2 and a fraction $1 \quad \epsilon$ of objects of type ω_1 to the mass of highest agent types $\theta > \hat{\theta}$, where $1 \quad G(\hat{\theta}) = F(\omega_2)+(1 \quad \epsilon)F(\omega_1)$, while it allocates the remaining fraction ϵ of objects ω_1 to agents types $\theta < \hat{\theta}$. Following the same procedure as above, we can compute the social welfare under M^{00}, i.e.,

$$\begin{aligned} W_{M^{00}} =& (0.5 \quad 0.25\epsilon)\left(\mathsf{E}_{\theta}\left[\theta \,j\, \theta > \hat{\theta}\right]\frac{(1 \quad \epsilon)\omega_1+\omega_2}{2 \quad \epsilon} \quad \hat{\theta}\left(\frac{(1 \quad \epsilon)\omega_1+\omega_2}{2 \quad \epsilon} \quad \frac{\epsilon\omega_1}{2+\epsilon}\right)\right) + \\ & (0.5+0.25\epsilon)\left(\mathsf{E}_{\theta}\left[\theta \,j\, \theta < \hat{\theta}\right]\frac{\epsilon\omega_1}{2+\epsilon}\right), \end{aligned}$$

which, for $\epsilon = 0.7$ and G being the arcsine distribution, equals

$$W_{M^{00}} = 0.136 \quad \omega_1 + 0.233 \quad \omega_2 > W_{M^{0}} > W_{M}.$$

Consequently, M, M^{0} are not optimal.

IV.4 The inclusion order of some classes of mechanisms

Fact IV.2. *The class of incentive-compatible monotone disjoint queue mechanisms is strictly smaller than the class of incentive-compatible disjoint queue mechanisms.*

Proof. By definition, every incentive-compatible monotone disjoint queue mechanisms is also a disjoint queue mechanisms. We will construct an incentive-compatible disjoint queue mechanism M which cannot be implemented by an incentive-compatible monotone disjoint queue mechanism. Consider a setting with $u(\theta,\omega)=\theta\omega$ and three object types $\omega_1=1$, $\omega_2=2$, $\omega_3=10$, such that $F(\omega_1)=\frac{1}{18}$, $F(\omega_2)=\frac{1}{18}$, $F(\omega_3)=\frac{8}{18}$. Let m be the median of the distribution of agent types G.

Mechanism M has two queues q_1,q_2: Agent types $\theta<m$ are sent to q_1. Objects of type ω_0,ω_2 are sent to q_1 at rates 0.5 $F(\omega_2)$ and $F(\omega_2)$, respectively. Thus, an agent who joins q_1 receives an object of expected type $X_M(\theta)=\frac{\omega_2 F(\omega_2)}{0.5}=\frac{1}{9}$. In q_2, the upper half of agents types θ m receive at random an object of either type ω_3 or ω_1, that is, $X_M(\theta)=\frac{\omega_1 F(\omega_1)+\omega_3 F(\omega_3)}{0.5}=9$.

We show that the outcome of M cannot be implemented by an incentive-compatible *monotone* disjoint queue mechanism $M^{\emptyset}$. This holds because, if such $M^{\emptyset}$ exists, then $X_{M^{\emptyset}}$ must take the form

$$X_{M^{\emptyset}}(\theta)=\begin{cases} X_{\mathrm{PAM}}(\theta), & \theta \not\in \bigcup_{i\in I}[\underline{\theta}_i,\bar{\theta}_i) \\ \frac{\int_{\underline{\theta}_i}^{\bar{\theta}_i} X_{\mathrm{PAM}}(s)\,\mathrm{d}G(s)}{G(\bar{\theta}_i)\quad G(\underline{\theta}_i)}, & \theta\in[\underline{\theta}_i,\bar{\theta}_i). \end{cases}$$

for a family of disjoint intervals $[\underline{\theta}_i,\bar{\theta}_i)$ indexed by $i\in I$.

Since $X_M(\theta)=X_{M^{\emptyset}}(\theta)$ and $X_M(\theta)=9<X_{\mathrm{PAM}}(\theta)$ for θ m, it must hold that $[m,\overline{\theta}]$ $[\underline{\theta}_i,\overline{\theta}_i]$ for some $i\in I$. Since $X_M(\theta)=\frac{1}{9}<9$ for all $\theta<m$, it follows that $\underline{\theta}_i=m$. This implies that

$$X_{M^{\emptyset}}(\theta)=\frac{\int_m^{\bar{\theta}} X_{\mathrm{PAM}}(s)\,\mathrm{d}G(s)}{G(\bar{\theta})\quad G(m)}=\frac{F(\omega_2)\omega_2+F(\omega_3)\omega_3}{0.5}=9.11>9=X_M(\theta)$$

holds for θ m, which is a contradiction. Consequently, no such $M^{\emptyset}$ can exist. □

Fact IV.3. *The class of incentive-compatible disjoint queue mechanisms is strictly smaller than the class of incentive-compatible mechanisms.*

Proof. We will construct an incentive-compatible mechanism M such that its outcome cannot be implemented by an incentive-compatible disjoint queue mechanism. Let $u(\theta,\omega)=\theta\omega$. There is a single object type ω_1 with $F(\omega_1)=\delta/2<1$. Mechanism M randomly assigns to each incoming agent an object of type ω_1 with probability $\delta\theta$; otherwise, it assigns her to the outside option ω_0. Thus, $X_M(\theta)=\delta\theta\omega_1$. By the envelope condition, the waiting cost should be $c_M(\theta)=0.5\delta\omega_1\theta^2$ for the mechanism to be incentive compatible.

We next show that no incentive-compatible disjoint queue mechanism $M^{\emptyset}$ can implement the same outcome as M. For contradiction, suppose that such $M^{\emptyset}$ exists and contains a finite number

of queues, k, as required by the definition. Since $M^{\emptyset}$ implements the same outcome as M, $c_{M^{\emptyset}}(\theta) = c_M(\theta)$ for all $\theta \in \Theta$. By the definition of disjoint queue mechanisms, for any $i \leq k$, all agents of types $\theta \in [\theta_i, \theta_{i+1})$ must be sent to queue q_i and incur the same waiting time, i.e., $c_{M^{\emptyset}}(\theta) = c_{M^{\emptyset}}(\theta^{\emptyset})$, for all $\theta, \theta^{\emptyset} \in [\theta_i, \theta_{i+1})$. This contradicts $c^{\emptyset}_M$ being strictly increasing. □

IV.5 Details for Example 2.2

In the setting described in Example 2.2 a mechanism M with reduced-form representation (X_M, c_M), where $X_M : \Theta \to [\omega_0, \omega_n]$ and $c_M : \Theta \to \mathbb{R}_+$, assigns to each agent with reported type $r := \mathsf{t}(\theta, a)$ an object of expected quality $X_M(r)$ at an expected waiting cost of $c_M(r)$. Given a mechanism M with (X_M, c_M), each agent picks an optimal action a and picks an optimal report r, according to the following optimization problem:

$$\max_{r \in \Theta} \left(\max_{a \in A} \left(X_M(r)\, \mathsf{t}(\theta, a) \right) - \mathsf{c}(a) \right) - c_M(r) \Big).$$

Let $u : [\omega_0, \omega_n] \times \Theta \to \mathbb{R}_+$ denote the payoff of an agent who receives an object of expected type ω and takes the optimal action; i.e., $u(\theta, \omega) = \max_{a \in A} \omega\, \mathsf{t}(\theta, a) - \mathsf{c}(a)$. Then, the above optimization problem can be equivalently written as follows:

$$\max_{r \in \Theta} u(\theta, X_M(r)) - c_M(r), \tag{IV.3}$$

which has the same form as the agent's maximization problem in the main setup.

The function u, albeit not being multiplicatively separable, satisfies the conditions required to apply Theorem 3.1. These are conditions i, ii and iii in Assumption 2.1.[7] To verify this, we first consider the simpler case in the example where $\mathsf{t}(\theta, a) = \theta + a$. Lemma 1 of Gershkov et al. (2021) implies that u is supermodular. Moreover, $u_1(\theta, \omega) = \omega$ is linear in ω, and thus it is convex in ω and bounded, since $|\omega| \leq \omega_n$. Furthermore, $u(\theta, \omega)$ is absolutely continuous in θ since $u(\theta, \omega)$ can be written as $u(\theta, \omega) = \max_{a \in A} \omega a - \mathsf{c}(a) + \omega\theta = u(0, \omega) + \int_0^{\theta} u_1(\theta, \omega)\, \mathrm{d}\theta$. Finally, u is absolutely continuous in ω since, for each fixed θ, $u(\theta, \cdot)$ it is a maximum over a finite number of functions that are linear in ω.

Given that the assumptions needed for Theorem 3.1 are satisfied, we can use the same analysis as in Appendix A to identify optimal mechanisms here. It remains to consider the second case described in the example where $\mathsf{c}(a) = \gamma \frac{a^b}{b}$ with $b > 1$ and $\gamma > 0$. Then,

$$u(\theta, \omega) = \frac{b-1}{b} \left(\frac{\theta^b \omega^b}{\gamma} \right)^{\frac{1}{b-1}}, \quad u_1(\theta, \omega) = \left(\frac{\theta^b \omega^b}{\gamma} \right)^{\frac{1}{b-1}} / \theta.$$

[7] Note that condition iv in Assumption 2.1 is not used to prove Theorem 3.1. In addition, *strict* supermodularity (i.e., condition i) is not required for Theorem 3.1; the proof requires only supermodularity. Strict supermodularity is used to prove the uniqueness properties discussed after Theorem 3.1.

Observe that u_1 is convex in ω, and u is absolutely continuous in each argument. By Lemma 1 of Gershkov et al. (2021), u is also supermodular. Thus, the conditions needed for Theorem 3.1 are satisfied and we can identify optimal mechanisms as in the previous case.

V Extensions

In Section V.1 we consider a more general notion of welfare which is a *weighted* average of the agents' payoffs. In Section V.2 we extend the main setup to allow for heterogeneity in the agent' waiting costs and also in the quality that they assign to each object. In Section V.3 we study scenarios where the agents have unobservable waiting costs.

V.1 Welfare as a weighted average of payoffs

We first show that Theorem 3.1 extends to the case where the planner defines welfare by a *weighted average* over the agents' payoffs rather than a simple average. Specifically, let $w : \Theta \; ! \; \mathsf{R}_+$ be an integrable function assigning a nonnegative weight to each agent type. The planner defines welfare under a mechanism M by

$$\int_0^{\bar{\theta}} w(\theta) \left(u(\theta, X_M(\theta)) \quad c_M(\theta)\right) \mathrm{d}G(\theta). \tag{V.1}$$

By (A.1), we can write weighted welfare as

$$\begin{aligned}\int_0^{\bar{\theta}} w(\theta) \left(u(\theta, X_M(\theta)) \quad c_M(\theta)\right) \mathrm{d}G(\theta) &= \int_0^{\bar{\theta}} w(\theta) \int_0^{\theta} u_1(\tau, X_M(\tau)) \,\mathrm{d}\tau \,\mathrm{d}G(\theta) \\ &= \int_0^{\bar{\theta}} u_1(\tau, X_M(\tau)) \int_\tau^{\bar{\theta}} w(\theta) \,\mathrm{d}G(\theta) \,\mathrm{d}\tau \\ &= \int_0^{\bar{\theta}} \mathsf{E}_{\tau \;\; G}[w(\tau) \; j \; \tau > \theta] \frac{1 \quad G(\theta)}{g(\theta)} u_1(\theta, X_M(\theta)) \,\mathrm{d}G(\theta). \end{aligned} \tag{V.2}$$

To adapt the proof of Theorem 3.1 to the setting with weighted welfare, we redefine the virtual welfare function as $\phi(\theta, x) = \mathsf{E}_{\tau \;\; G}[w(\tau) \; j \; \tau > \theta] \frac{1 \quad G(\theta)}{g(\theta)} u_1(\theta, x)$. Observe that (V.2) is a convex functional of X_M, for the same reasons as in the original proof. The rest of the proof applies identically. Also, the results of Section 4 hold identically for the above generalization, as well as the main result in our comparative statics analysis, Theorem 5.1. The only change in the proofs is that welfare will be defined by (V.1).

V.2 Agents may disagree on the objects' qualities

We next consider an extension of our model where agents belong to one of l groups indexed by $j \; 2 \; [l]$. The group of an agent is an observable characteristic. Agents from each group $j \; 2 \; [l]$

arrive at a rate a_j, with $\sum_{j=1}^{l} a_j = 1$. An agent's type is her private information, and is distributed according to $G^{(j)}$ for agents in group j. Agents also have group-specific utility functions: the utility function of an agent in group j is $u^{(j)}(\theta, \omega)$. Each function $u^{(j)}$ satisfies Assumption 2.1.

Extend the definition of mechanisms from Section 2 to this setting so that a mechanism M can condition on an agent's group when assigning her an allocation timeline. Thus, we denote a mechanism here by $M = (M_j)_{j \in [l]}$, where M_j denotes a mechanism in our baseline setup that allocates objects to agents in group j. Let $\overline{X}$ be the set of interim allocation rules of all individually rational and incentive-compatible mechanisms M in this extended setting. As in Appendix A.2, without loss of generality we can focus on the reduced-form $(\mathbf{X}_M, \mathbf{c}_M)$ of a mechanism M, where $\mathbf{X}_M = (X_j)_{j \in [l]}$ and $\mathbf{c}_M = (c_j)_{j \in [l]}$. For each $j \in [l]$, $X_j(\theta)$ and $c_j(\theta)$ respectively give the expected object type and expected waiting cost for agents of type θ in group j. We next prove an analog of Theorem 4.1 for this setting.

Proposition V.1. *In a setting with l groups of agents, the optimal mechanism $M = (M_j)_{j \in [l]}$ can be implemented by l FCFS waitlists with deferrals, one per agent group j, paired with an information disclosure policy μ_j that pools adjacent object types.*

Proof. We break the proof into three steps. First, we show the existence of an optimal mechanism. Second, conditional on the existence of an optimal mechanism, we show that each M_j must be a monotone disjoint queue mechanism. Third, we apply Theorem 4.1 for each M_j.

Similar to Lemma A.2, the planner's objective can be written as

$$O := \sup_{\mathbf{X} \in \overline{X}} \sum_{j=1}^{l} a_j \int_0^{\bar{\theta}} \left(u^{(j)}(\theta, X_j(\theta))\lambda_E + \phi^{(j)}(\theta, X_j(\theta))\lambda_W \right) \mathrm{d}G^{(j)}(\theta), \tag{V.3}$$

where $\phi^{(j)}(\theta, x) = \frac{1 - G^{(j)}(\theta)}{g^{(j)}(\theta)} u_1^{(j)}(\theta, x)$.

Observe that a feasible solution $\mathbf{X} = (X_j)_{j \in [l]} \in \overline{X}$ exists. Moreover, the value of (V.3) is bounded by the sum of optimal objectives for each group j when they receive all of the objects. Consequently, an optimal mechanism M exists. We next show that there is an optimal mechanism M such that, for each j, M_j takes the form of a monotone disjoint queue mechanism.

Let M be an optimal mechanism for (V.3). For each group j, we denote the total rate of objects of type ω_i that M assigns to agents in group j by $b_j(\omega_i)$; naturally, $\sum_{j=1}^{l} b_j(\omega_i) \quad F(\omega_i)$ must hold. Also, for all agent groups j, let $X^{(j)}$ denote the set of all incentive-compatible and individually rational mechanisms M_j which, for all object types ω_i, allocates objects of type ω_i to the agents in group j at a rate of at most $b_j(\omega_i)$. For each such mechanism M_j, Theorem 3.1 implies that there is a monotone disjoint queue mechanism $\widetilde{M}_j$ (with corresponding interim allocation rule

$\widetilde{X}_j$) that achieves the following objective for group j:

$$O_j(b_j) := \sup_{X_j \in \mathcal{X}^{(j)}} \int_0^{\bar{\theta}} \left(u^{(j)}(\theta, X_j(\theta))\lambda_E + \phi^{(j)}(\theta, X_j(\theta))\lambda_W \right) \mathrm{d}G^{(j)}(\theta). \tag{V.4}$$

On the other hand, the optimal value in (V.3) equals:

$$\begin{aligned} O &= \sum_{j=1}^{l} \alpha_j \int_0^{\bar{\theta}} \left(u^{(j)}(\theta, X_j(\theta))\lambda_E + \phi^{(j)}(\theta, X_j(\theta))\lambda_W \right) \mathrm{d}G^{(j)}(\theta) \\ &= \sum_{j=1}^{l} \alpha_j O_j(b_j) \\ &= \sum_{j=1}^{l} \alpha_j \int_0^{\bar{\theta}} \left(u^{(j)}(\theta, \widetilde{X}_j(\theta))\lambda_E + \phi^{(j)}(\theta, \widetilde{X}_j(\theta))\lambda_W \right) \mathrm{d}G^{(j)}(\theta), \end{aligned}$$

where the first equality follows from the optimality of $\mathcal{M}$, the second equality follows from the optimality of each $\mathcal{M}_j$, and the third from the optimality of $\widetilde{\mathcal{M}}_j$ for each j. Thus, $\widetilde{\mathcal{M}}$ achieves the optimal value O.

For the third step, consider a group $j \in [l]$. By Theorem 4.1, there is an information disclosure policy μ_j such that $\mathcal{M}_j$ can be implemented via a single FCFS waitlist with deferrals Q_j paired with μ_j. Thus, the system of l queues $Q_1, \ldots, Q_l$ implements $\mathcal{M} = (\mathcal{M}_j)_{j \in [l]}$. The second step shows that $\mathcal{M}$ attains the second-best objective value. Hence, so does this system of l queues. □

V.3 Unobservable waiting cost functions

We consider an extension of our baseline model to a setting where the *type* of an agent is characterized by two parameters: her *valuation for quality* θ and her *impatience level* η. The payoff of the agent if she gets allocated an object of quality ω after waiting t units of time is $\theta\omega - \eta t$. We call $r = \theta/\eta$ the agent's *rate of substitution between quality and waiting time*. The pair (θ, η) is distributed according to a continuous joint CDF H. Let H_r be the CDF of the marginal distribution induced over r by H. We denote the PDF of H_r by h_r. Also, denote the support of r by $[0, \bar{r}]$.

A mechanism is defined as in our baseline setup, with the difference that an agent who reports a type (θ, η) is assigned an allocation timeline that depends on both θ, η. Thus, the expected object type assigned by the mechanism to the agent now depends on both θ and η; we denote this expectation by $\bar{X}(\theta, \eta)$. Similarly, we denote the expected waiting cost of the agent by $\bar{c}(\theta, \eta)$. We call $(\bar{X}, \bar{c})$ the *reduced form* of the mechanism. The interim incentive-compatibility constraint then is

$$\theta\, \bar{X}(\theta, \eta) - \eta\, \bar{c}(\theta, \eta) \geq \theta\, \bar{X}(\theta', \eta') - \eta\, \bar{c}(\theta', \eta')$$

for all types (θ, η) and (θ', η'). The interim individual rationality constraint is

$$\theta\, \bar{X}(\theta, \eta) \quad \eta\, \bar{c}(\theta, \eta) \quad 0.$$

The planner's objective remains maximizing a weighted sum of welfare and efficiency.

Our results in Sections 3 and 4 translate to this setting. We first show this for Theorem 3.1. The high-level idea is that a mechanism can at most elicit the substitution rate of the agents, and thus can be represented by a one-dimensional reduced-form mechanism $(X(r), t(r))$, without any loss in the objective. This makes the analysis from the main setup applicable, as follows.

The description of monotone disjoint queue mechanisms remains the same as in the main setup, with the difference that an agent's type θ in that description is replaced with her substitution rate r in here. In particular, a monotone disjoint queue mechanism is defined by a strictly increasing sequence of reals $r_0 = 0, r_1, \ldots, r_k, r_{k+1} = \bar{r}$ such that an agent who reports a substitution rate r 2 $[r_i, r_{i+1})$ for i 0 is assigned upon arrival to the queue q_i. The rest of the definition (the allocation of objects across and within queues) remains the same as in the main setup. The main technicality to apply the analysis from the main setup is proving the following proposition.

Proposition V.2. *For an incentive-compatible and individually rational mechanism $\bar{M}$ with reduced form $(\bar{X}(\theta, \eta), \bar{c}(\theta, \eta))$, there exists a mechanism $\hat{M}$ with reduced form $(\hat{X}(\theta, \eta), \hat{t}(\theta, \eta))$ such that $\hat{M}$ achieves the same objective as $\bar{M}$, and $\hat{X}(\theta, \eta) = \hat{X}(\theta', \eta')$ when $\theta/\eta = \theta'/\eta'$. Moreover, the value of the objective achieved by $\hat{M}$ is*

$$\int_0^{\bar{r}} w(r) \left(\lambda_W \frac{1 \quad H_r(r)}{h_r(r)} X(r) + \lambda_E\, r X(r) \right) \mathrm{d}H_r(r), \tag{V.5}$$

where $w(r) = \mathsf{E}_\eta[\eta \,j\, r]$ and $X(r) = \hat{X}(r\eta, \eta)$ for all r, η.

Proof. The proof mirrors the arguments in Theorem 8 of Dworczak et al. (2021). Since $(\bar{X}, \bar{c})$ is incentive-compatible, the following inequalities hold for any two agent types (θ, η) and (θ', η'):

$$\theta\, \bar{X}(\theta, \eta) \quad \eta\, \bar{c}(\theta, \eta) \quad \theta\, \bar{X}(\theta', \eta') \quad \eta\, \bar{c}(\theta', \eta'),$$

$$\theta'\, \bar{X}(\theta', \eta') \quad \eta'\, \bar{c}(\theta', \eta') \quad \theta'\, \bar{X}(\theta, \eta) \quad \eta'\, \bar{c}(\theta, \eta).$$

These two inequalities together imply that

$$(\bar{X}(\theta, \eta) \quad \bar{X}(\theta', \eta')) \begin{pmatrix} \theta & \theta' \\ \eta & \eta' \end{pmatrix} \quad 0,$$

that is, $\bar{X}$ is non-decreasing in θ/η. This implies that there are non-decreasing functions X, t :

$[0, \bar{r}]$! $\overline{\Omega}$ such that almost everywhere

$$\bar{X}(\theta, \eta) = X(\theta/\eta) \quad \text{and} \quad \bar{c}(\theta, \eta) = t(\theta/\eta).$$

This is Lemma 3 in Dworczak et al. (2021).

Analogous to Myerson (1981), it follows that (X, t) is incentive-compatible. It remains to compute the objective achieved by (X, t).

$$\begin{aligned}
&\int_\theta \int_\eta \left(\lambda_W \left(\theta\, \bar{X}(\theta, \eta) \quad \eta\, \bar{c}(\theta, \eta)\right) + \lambda_E \theta \bar{X}(\theta, \eta)\right) \mathrm{d}H(\theta, \eta) \\
&= \int_\theta \int_\eta \left(\lambda_W \left(\theta\, X(\theta/\eta) \quad \eta\, t(\theta/\eta)\right) + \lambda_E \theta X(\theta/\eta)\right) \mathrm{d}H(\theta, \eta) \\
&= \int_0^{\bar{r}} \mathsf{E}_\eta[\eta \,|\, r] \left((\lambda_W + \lambda_E) r\, X(r) \quad \lambda_W t(r)\right) \mathrm{d}H_r(\theta) \\
&= \int_0^{\bar{r}} w(r) \left((\lambda_W + \lambda_E) r\, X(r) \quad \lambda_W t(r)\right) \mathrm{d}H_r(\theta) \\
&= \int_0^{\bar{r}} w(r) \left(\lambda_W \frac{1 \quad H_r(r)}{h_r(r)} X(r) + \lambda_E\, r X(r)\right) \mathrm{d}H_r(r),
\end{aligned}$$

where the last step follows from the analysis in Online Appendix V.1. □

The above proposition reduces the two-dimensional reduced form problem to a one-dimensional problem in our main setup. The analysis in Online Appendix V.1 applies to this one-dimensional reduced form and implies that the optimal mechanism is a monotone disjoint queue mechanism; i.e., a counterpart for Theorem 3.1. Similarly, Theorem 4.1 holds here as well by the same proof, but for the substitution rate parameter r replacing the one-dimensional type parameter θ there.