Inefficiencies and Congestion in Waiting Lists

Ingredients in “typical” waiting lists

Arrivals of agents and items

- Items of different types
- Agents have private preferences - private type .

  Uility for item j given item waiting times $w = (w_1, w_2, ..)$: $\mathrm{u}_\theta(\mathrm{j}, \mathrm{w})$

Today: sources of inefficiency

- Randomness in arrivals of agents and items
  - Ashlagi, Qian, Leshno, Saberi (2022)
- Perishable objects
  - Ashlagi, Jagadeesan and Qian (2024)

A model with random arrivals and quasi-linear utilities

Items: Arrive according to Poisson process, total rate $\mu = 1$

- Finite types: $J = \{1,2,\ldots,J\}$
- With probability μ_j arriving item is of type j

Agents: Arrive according to Poisson process with total rate λ

- Agent type $\theta \in \Theta$, drawn i.i.d. according to distribution F
- Possibly uncountably many or finitely many types

Quasi-Linear Utility: Type θ agent who is assigned j and waits w has utility:

$$\mathrm{u}_\theta(j,w) = v(\theta,j) - c(w)$$

- Agents can leave immediately (balk) to obtain utility $v_\theta(\phi) = 0$
- Match values are private information
- $v(\theta,j)$ is bounded; $c(\cdot)$ is smooth, strictly increasing and convex or concave

Price Discovery in Waiting Lists with Random Arrivals

Question: what is (allocative) inefficiency due to fluctuating “prices”?

- Natural price discovery process
  - Tâtonnement processes – price increases with demand (agents join queue), decreases with supply (items arrive)
- Key distinction: prices fluctuate over time
  - Prices are not specified, but learned
  - Changes with each random arrival of agent or item
  - Prices discovery never stops

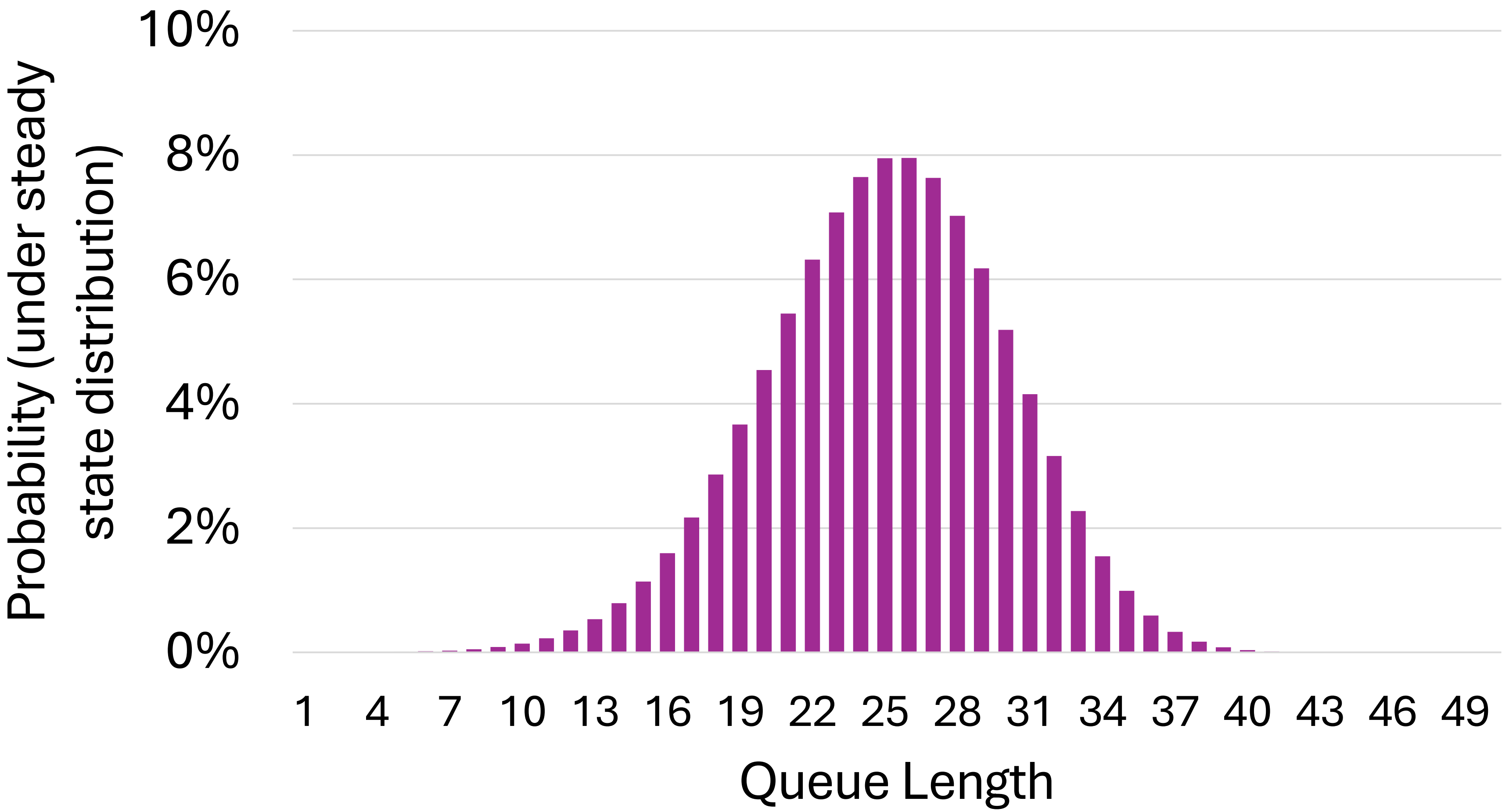
Example – One Item

- Single item, arrives at Poisson rate 1
- Agents arrive at Poisson rate 2
  - An agent’s value for the item is $v \sim U[0,1]$ i.i.d.
  - Agents can join the queue, or leave immediately
  - Quasilinear utility

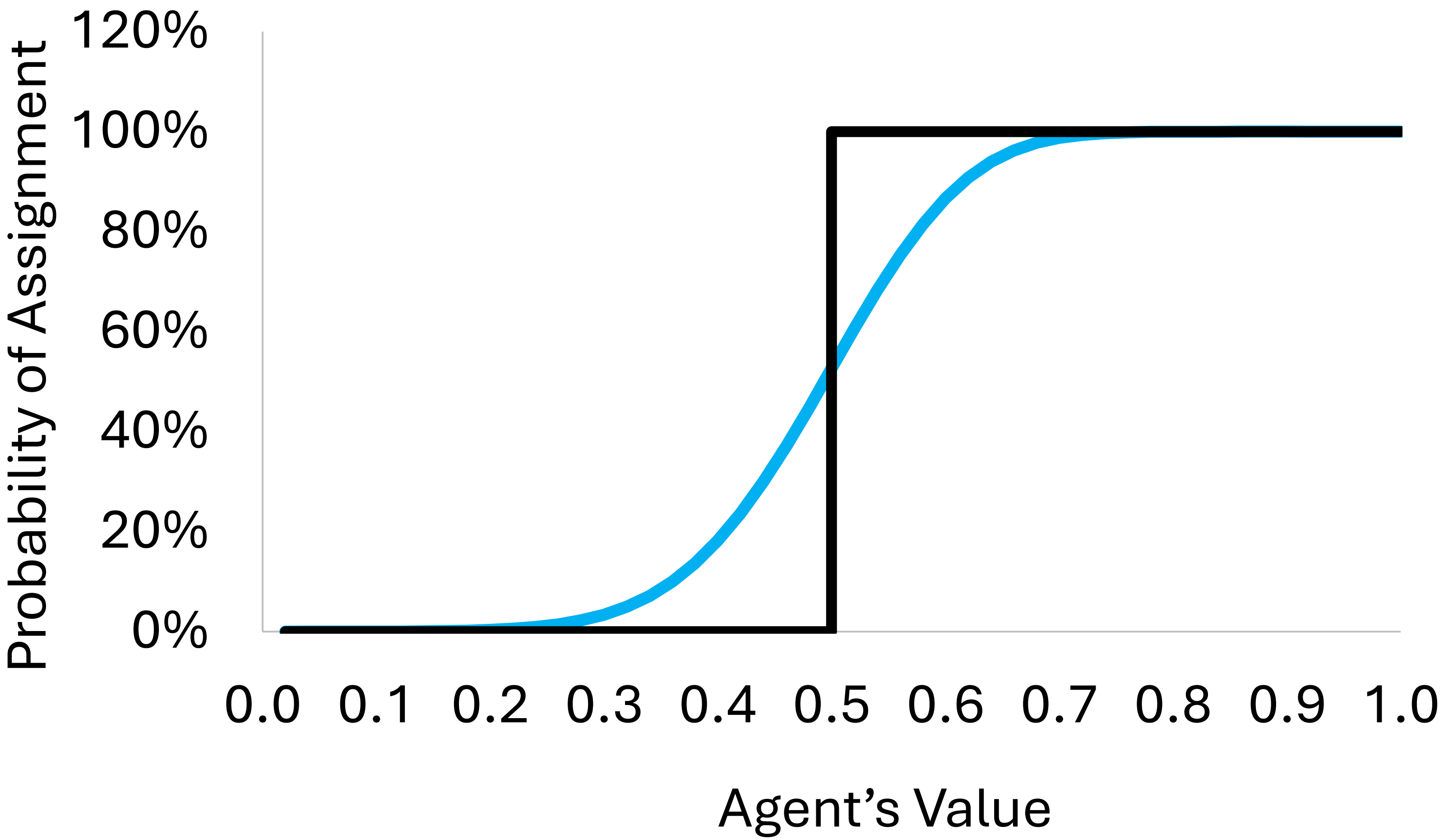
$$v - 0.02 \cdot w$$

- Offline benchmark:
  - Collect all items and agents that arrive until (large) time T
  - Optimal to set a price of $1/2$, assigning agents with $v \in [\frac{1}{2}, 1]$

Example – One Item



Example – One Item



Price Discovery in Waiting Lists

Question: what is allocative inefficiency due to fluctuating “prices”?

- Result: Loss from price fluctuations is bounded by the *step size*
  - Bound is tight
  - Conditions for when loss is negligible

- Methodology: “Price adaptation” as a stochastic gradient decent (SGD)
  - Duality, Lyapunov functions

Literature

Dynamic matching mechanisms:

- Leshno 2017, Baccara, Lee and Yariv 2018, Bloch Cantala 2017, Su and Zenios 2004, Arnosti and Shi 2017, Loertscher Muir Taylor 2020,

.

Convergence of tâtonnement processes using gradient descent:

- In markets with multiple goods Cheung, Cole and Devanur, 2019, Cheung, Cole and Tao, 2018, Cole and Fleischer, 2008, Uzawa, 1960, Correa and Stier-Moses, 2010, Powell and Sheffi, 1982.

(Centralized) Dynamic matching:

- Busic and Meyn 2014, Gurvich and Ward, 2014, Nazari and Stolyar 2016, Kerimov, Ashlagi and Gurvich 2021a,b

Assignments and Allocative Efficiency

- Assignments $\boldsymbol{\eta}$

  Let $\eta_t \in J_\emptyset$ denote the item assigned to agent who arrived at $t \in \mathcal{A}_T$, where $\mathcal{A}_T$ are the arrival epochs in which an agent arrives

- Allocative efficiency

$$W(\eta) = \liminf_{T \to \infty} \frac{1}{|\mathcal{A}_T|} \sum_{t \in \mathcal{A}_T} v(\theta_t, \eta_t)$$

  That is, average assigned value per agent

- Optimal allocative efficiency

$$W^{OPT} = \mathbb{E}\left[\sup_{\eta} W(\eta)\right]$$

  - Restricting attention to assignments η that satisfy a no-Ponzi condition

The Waiting List Mechanism

- Separate queue for each item $j \in J$
  - First Come First Served (FCFS) assignment policy
  - Agents who join a queue wait until assigned (no reneging)

- Choice of agent θ who observes $\boldsymbol{q}$:

$$a(\theta, \mathbf{q}) = \underset{j \in \mathcal{J} \cup \{\emptyset\}}{\operatorname{argmax}} \left\{ v(\theta, j) - \mathbb{E}[c(w_j) | \mathbf{q}] \right\}$$

  - Observes all queue lengths $\boldsymbol{q} = \left(q_1, \ldots, q_J\right)$
  - Can join any queue, or leave unassigned

The Waiting List Mechanism

- Separate queue for each item $j \in J$
  - First Come First Served (FCFS) assignment policy
  - Agents who join a queue wait until assigned (no reneging)

- Choice of agent θ who observes $\boldsymbol{q}$:

$$a(\theta, \mathbf{q}) = \underset{j \in \mathcal{J} \cup \{\emptyset\}}{\operatorname{argmax}} \left\{ v(\theta, j) - p_j(\mathbf{q}) \right\}$$

  - Observes state-dependent prices:

$$p_j(\boldsymbol{q}) = p_j(q_j) = \mathbb{E}[c(w_j) | \, q_j]$$

Stochastic price adaptation

- Prices increase and decrease upon arrival and allocation

- Allocative efficiency is the expected match value under the steady state distribution

- When there are >2 items, the steady state distribution is not tractable

The Waiting List Mechanism

- The **adjustment size Δ** is the maximal change from one arrival:

$$\Delta = \max_{j \in \mathcal{J}} \max_{1 \le q \le q_{\max}} \{p_j(q) - p_j(q-1)\}$$

  - For linear waiting costs, $c(w) = c \cdot w$, adjustment size is the maximal cost of waiting for one item's arrival:

$$\Delta = {}^{c}/_{\mu_{min}}$$

- Denote the expected allocative efficiency under the waiting list

$$W^{WL} = \mathbb{E}[W(\eta^{WL})]$$

Bounding Allocative Efficiency

Theorem:

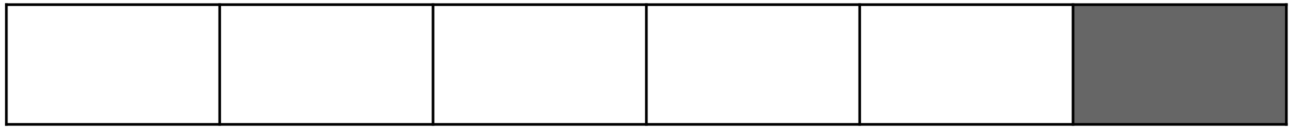
Allocative efficiency under the waiting list is bounded by

$$W^{WL} \geq W^{OPT} - \frac{\lambda + 2}{2\lambda} \Delta$$

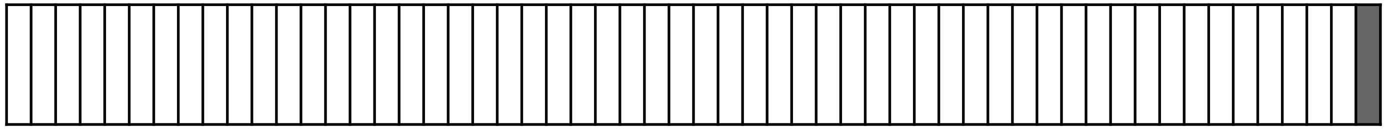
=> The allocative efficiency loss is bounded by the cost of waiting for one item arrival. High loss if an item arrives infrequently, low loss if the item arrives frequently

Main Result: Intuition

- **Suppose $p^* =$ cost of waiting six months**
  - If an item arrives monthly, corresponding queue length is 5
  - Each arrival significantly changes the price



  - If an item arrives daily, corresponding queue length is 180
  - Each arrival slightly changes the price



Relation to Static Assignment

Lemma: $W^{OPT} = W^*$

W^* is the optimal allocative efficiency in the corresponding static assignment problem:

$$
\begin{aligned}
W^* = \max_{\{x_{\theta j}\}_{\theta \in \Theta, j \in \mathcal{J}}} \quad & \sum_{j \in \mathcal{J}} \int_{\Theta} x_{\theta j}\, v(\theta, j)\, dF(\theta) & \\
\text{subject to} \quad & \sum_{j \in \mathcal{J}} x_{\theta j} \leq 1,\ x_{\theta j} \in [0,1] & \forall \theta \in \Theta \\
& \int_{\Theta} \lambda x_{\theta j}\, dF(\theta) \leq \mu_j & \forall j \in \mathcal{J}
\end{aligned}
$$

Duality for the Static Assignment

Lemma (Monge-Kantorovich duality):

$$\min_{\boldsymbol{p}\geq\boldsymbol{0}} h(\boldsymbol{p}) = W^*$$

for

$$h(p) = \int_{\Theta} \max_{j\in J\cup\{\emptyset\}}\left[v(\theta,j) - p_j\right] + \frac{1}{\lambda}\sum_{j\in J}\mu_j p_j$$

Relation to Stochastic Gradient Descent

The expected adjustment is

$$\mathbb{E}[q_{j,t+1} - q_{j,t}] = \frac{\lambda}{1+\lambda} \int_{\Theta} \mathbf{1}_{\{a(\theta,\boldsymbol{q}_t)=j\}} dF(\theta) - \frac{1}{1+\lambda}\mu_j$$

which is a sub-gradient of the dual objective

$$h(\boldsymbol{p}) = \int_{\Theta} \max_{j \in \mathcal{J} \cup \{\emptyset\}} [v(\theta, j) - p_j]\, dF(\theta) + \frac{1}{\lambda} \sum_{j \in \mathcal{J}} \mu_j p_j$$

- That is, the expected step is in a gradient descent direction
- But unlike when SGD is used for optimization, step size is fixed and does not shrink to 0

Proof Idea

- Define a Lyapunov function $L(q)$ such that $\nabla L(q) = p(q)$

- Using the dual objective we decompose and bound the value generated from an arrival in state q_t:

$$\begin{aligned}\mathbb{E}[v(\theta_t, a(\theta_t, \mathbf{q}_t))|\mathbf{q}_t] \geq\ & \frac{\lambda}{\lambda+1} W^* \\ & \underbrace{- \frac{1}{\mu_{\min} \cdot \Delta} \left(L(\mathbf{q}_t) - \mathbb{E}[L(\mathbf{q}_{t+1})|\mathbf{q}_t]\right)}_{\text{(I) Change in Potential}} \\ & \underbrace{- \frac{2+\lambda}{2(1+\lambda)} \Delta}_{\text{(II) Loss}}\end{aligned}$$

Proof Idea

- Decompose the value generated from an arrival in state q_t:

$$\begin{aligned}\mathbb{E}[v(\theta_t, a(\theta_t, \mathbf{q}_t))|\mathbf{q}_t] \geq\ & \frac{\lambda}{\lambda+1}W^* \\ & - \underbrace{\frac{1}{\mu_{\min}\cdot\Delta}\left(L(\mathbf{q}_t) - \mathbb{E}[L(\mathbf{q}_{t+1})|\mathbf{q}_t]\right)}_{\text{(I) Change in Potential}} \\ & - \underbrace{\frac{2+\lambda}{2(1+\lambda)}\Delta}_{\text{(II) Loss}}\end{aligned}$$

○ *state independent*

○ *state dependent*

Example of High Loss (of order Δ)

- Agents $\Theta = J$, each agent only wants one corresponding item

$$v(\theta, j) = \mathbf{1}_{\{\theta = j\}}$$

- Identical arrival rates of items and corresponding agents

- Loss is close to Δ
  - Queue lengths follow an unbiased reflected random walk
  - Queue lengths $q_j = 0,1,2,\dots,1/\Delta = \mu_j/c$ equally likely in steady state
  - Loss when an agent arrives and price is too high, |
    i.e., queue length hits its boundary
  - Probability of hitting the boundary is roughly ${}^{1}/_{1/\Delta}$.

When is the Loss Small?

Theorem: Assume an economy with finitely many agent types, linear waiting costs $c(w) = c \cdot w$, and a unique market clearing price.

Then there exist $\alpha, \beta, c_0 > 0$ such that for any $c < c_0$

$$W^{WL} \geq W^{OPT} - \beta e^{-\alpha/\Delta}$$

- Remark: generically, an economy with finitely many agents has a unique market clearing

Intuition: If the dual is unique, no loss within a neighborhood of p^*
Biased random walk towards p^*

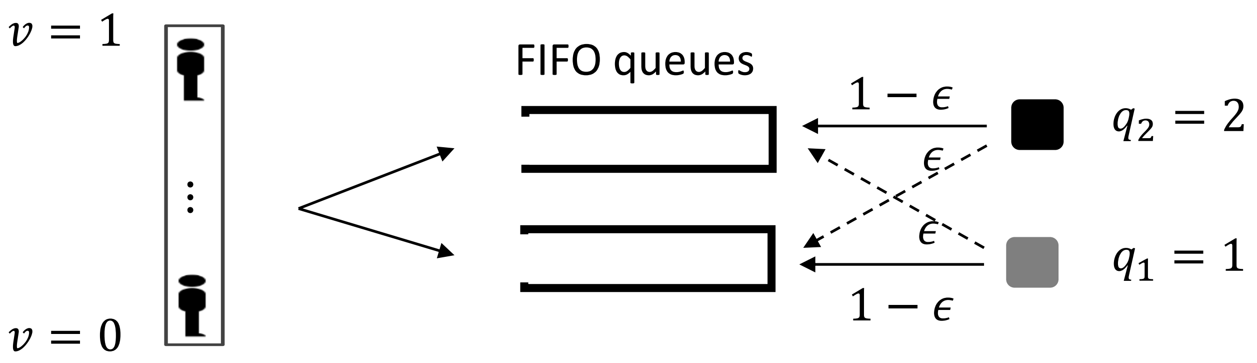
Waiting times, allocation and welfare

- When payoffs are quasi-linear with waiting times the outcome is “almost” allocative efficient

- It is not welfare maximizing as agents “waste” time in queues. To maximize welfare, some pooling or randomization is necessary

Example: Why is pooling necessary for welfare?

- Agents arrive at rate 1, types drawn from $\mathrm{Uniform}[0,1]$
- Objects arrive at rate 1, qualities drawn uniformly from $\{1,2\}$
- $u(v,q) = vq$
- A disjoint queue mechanism:



- $\epsilon = 0 \Rightarrow$ No pooling (one queue per object)
- $\epsilon = \frac{1}{2} \Rightarrow$ Complete pooling
- The agent with $v = 0.5$ should be indifferent…

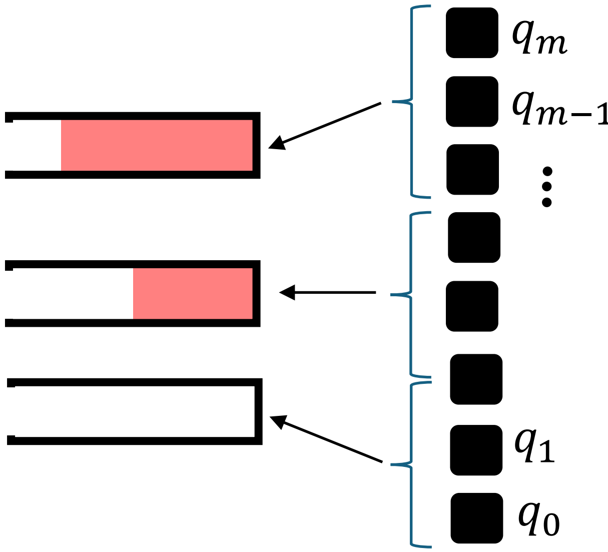
Waiting times, allocation and welfare

- When payoffs are quasi-linear with waiting times the outcome is “almost” allocative efficient

- It is not welfare maximizing agents “waste” time in queues. To maximize welfare, some pooling or randomization is necessary

- When objects have common qualities:
  - There is a monotone disjoint queueing mechanism (system of queues with pooled adjacent types) which generate “almost” optimal welfare (Ashlagi, Monachou, Nikzad, ReStud 2023)
    - Agents pick one queue and cannot decline an object

But an open question in multidimensional settings…



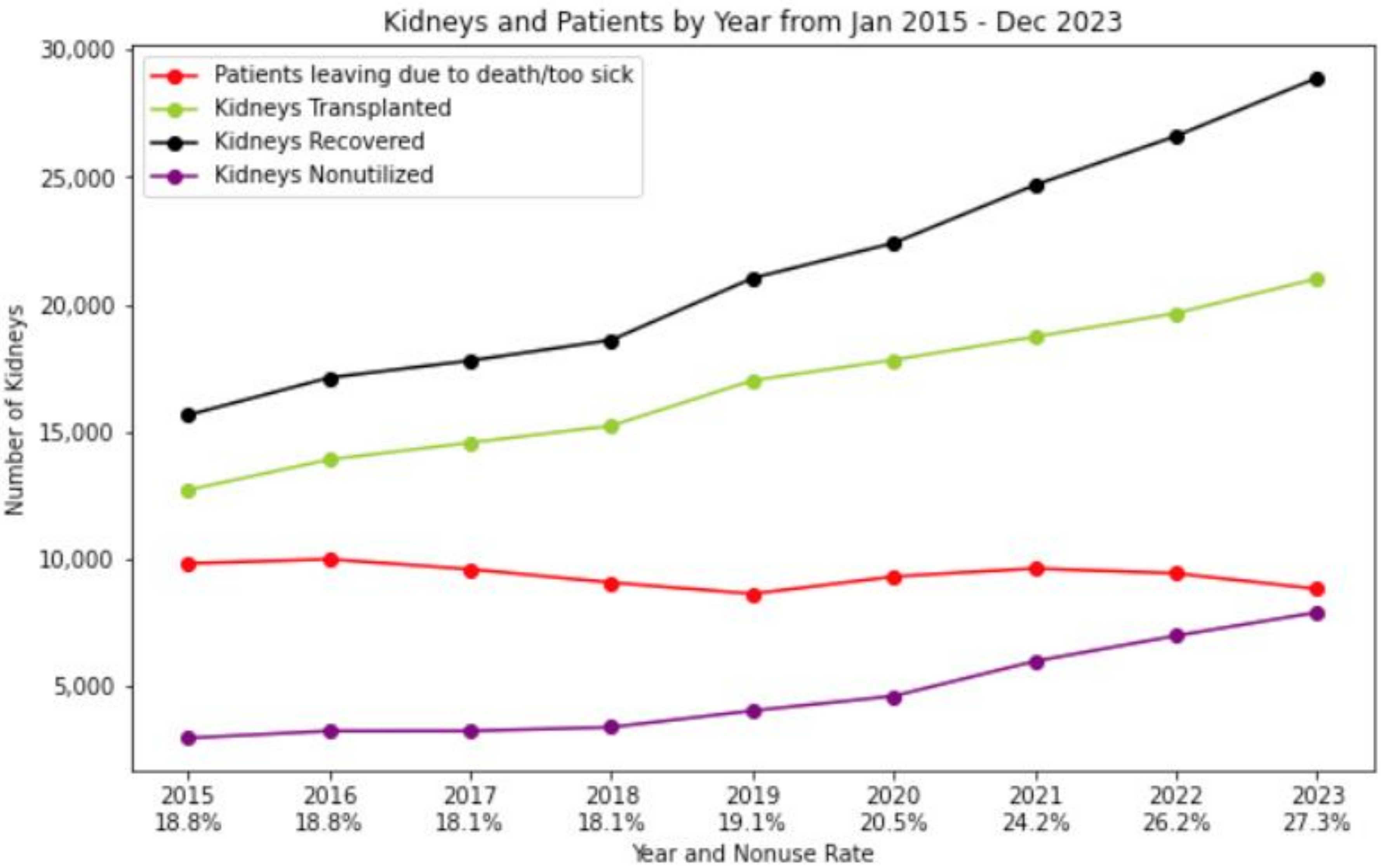
Today: sources of inefficiency

- Randomness in arrivals of agents and items
  - Ashlagi, Qian, Leshno, Saberi (2022)

- Perishable objects
  - Ashlagi, Jagadeesan and Qian (2024)

Congestion in Waiting Lists and Organ Allocation

Waiting list for kidneys from deceased donors in the US (2015-2023)



~90k patients on the waiting list today

The paper

Organ waiting lists can become *congested*, leading to discard of valuable organs

- Patients near the top of the list may decline to accept low-quality organs
- Patients further down the list might very well accept one

- Friction: limited number of patients can consider an organ before it expires
  - kidneys accrue excess cold ischemic time

- so lower quality organs may expire before being offered to patients who would accept them (despite there potentially being many such patients)

- Goal: formalize this force and investigate implications for welfare, discuss moral hazard, and design

A fluid model with agents' departures

- continuum of patients arrive at rate *p*
- safe organs arrive at rate *sp* (*s* < 1) and risky organs arrive at rate *rp*
- organs offered sequentially to patients in descending order of waiting time

- **patients of type $\boldsymbol{\theta}$ leave w/o a match at rate $\boldsymbol{\delta(\theta)}$**
- patients have expected utility preferences and type θ get utility:

$$\boldsymbol{u_{patient} = E[1(\text{get safe organ}) + \nu(\theta)1(\text{get risky organ})], \quad \nu(\theta) \leq 1}$$

- patients can decline risky organs and hold out for safe organs

- **organs expire if declined by $\boldsymbol{\epsilon}$ mass of patients** ("cold ischemic time" limit)
  - in this case (or if no one wants them) then organs are discarded

Literature

deceased donor allocation

Su & Zenios (2004, 2006), Bertsimas, Farias & Trichakis (2013), Ata, Friedewald & Randa (2020), Agarwal, Ashlagi, Waldinger, Somaini & Rees (2021),Agarwal, Hodgson & Somaini (2021), Kang, Koren, Monachou, Ashlagi (2021), **Shi & Yin (2022),** Chan & Roth (2023), Sweat (2023), Bae (2024)

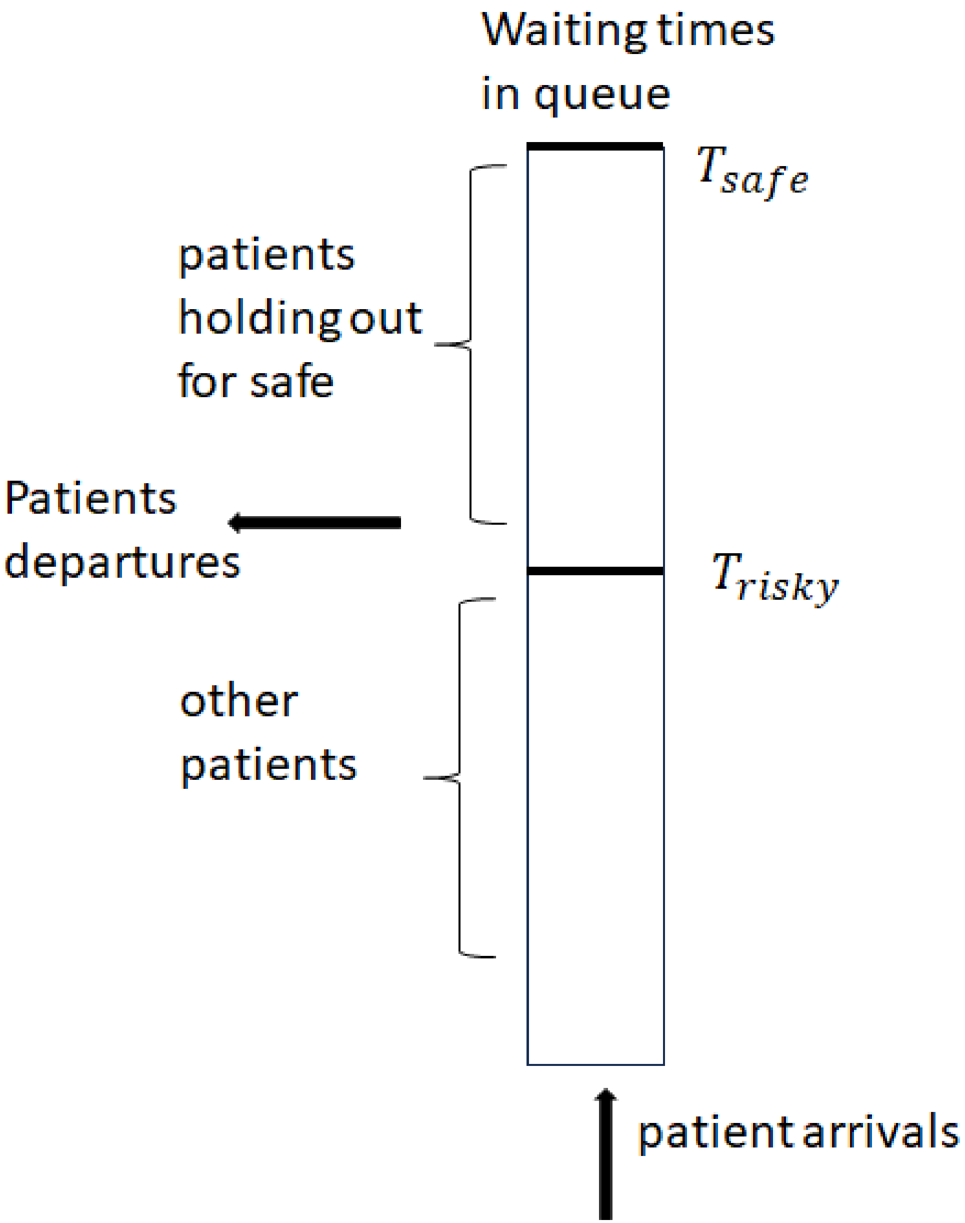
wait list design

Naor (1968), Hassin & Haviv (2012), Bloch & Cantala (2017), Baccara, Lee & Yariv (2018), Arnosti & Shi (2020), Leshno (2020), Ashlagi, Monachou & Nikzad (2023), Nikzad and Strack (2023), Shmaya & Scarsini (2024), Che and Tercieux (2024)

waiting list w/perishable objects

Castro, Ma, Nazerzadeh & Yan (2020)

Structure of steady states



- T_{safe}, T_{risky} are equilibrium objects
- patients of type θ who are offered a risky kidney will decline it and hold out for a safe kidney if
- $T_{safe} - T_{risky} \leq \frac{-log v(\theta)}{\delta(\theta)} = W_{v}(\theta)$

Existence of steady states

- assumption: distribution of $W_\nu(\theta)$ is absolutely continuous and W_ν and δ are bounded from above

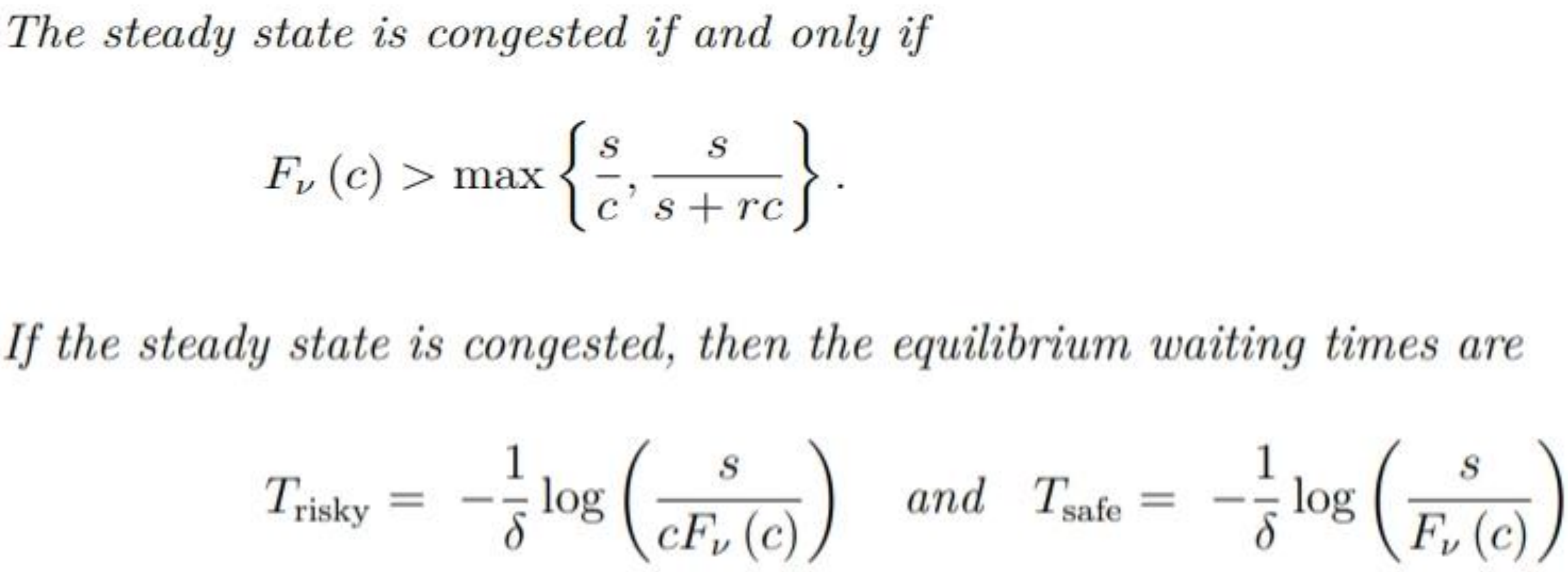
Theorem:

There exists a steady-state joint distribution of types and waiting times

- key equilibrium objects (prices): waiting times T_{safe} and T_{risky}
- patients w/wait time T_{safe} are offered safe organs and accept it
- patients w/wait time ≥ T_{risky} are offered risky organs
  - types with $W_\nu(\theta) < T_{safe} - T_{risky}$ accept
  - Types with $W_\nu(\theta) > T_{safe} - T_{risky}$ decline and wait for safe organs

T_{safe} and T_{risky} are sufficient statistics for all types' expected utilities

Homogenous departure rates

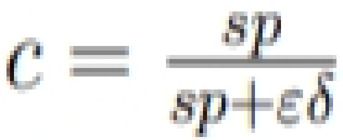


The steady state is congested if and only if

$$F_\nu(c) > \max\left\{\frac{s}{c}, \frac{s}{s+rc}\right\}.$$

If the steady state is congested, then the equilibrium waiting times are

$$T_{\text{risky}} = -\frac{1}{\delta}\log\left(\frac{s}{cF_\nu(c)}\right) \quad \textit{and} \quad T_{\text{safe}} = -\frac{1}{\delta}\log\left(\frac{s}{F_\nu(c)}\right)$$



$$c = \frac{sp}{sp+\varepsilon\delta}$$

c is a “degree of congestion “

Inefficiency and discard can arise even without expiration
(Shi and Yin 2022)

- waiting times T_{safe} and T_{risky} act as prices but don't induce allocative efficiency

even when $\epsilon = \infty$:

- Safe organs are "overdemanded" by patients with low $\nu(\theta)$ (low value for a risky organs)
  - Such patients "bet" on being offered a safe organ
- But patients who wait for a safe organ may die before getting it
  - Despite that they would have been better off with a risk organ

- So the allocation is ex post inefficient, discard arises, and patients die unnecessarily

*Notes:

- waste can be eliminated using lotteries upon entry
- If utilities are instead quasi linear with waiting time, the allocation is assortative

Congestion

- In standard queuing systems: $T_{risky} > 0$ ⟺ there is excess demand for risky organs

Definition

The system is **congested** if $T_{risky} > 0$ but some risky organs are discarded

The system is congested when ϵ is small enough:

Proposition

There exists $\epsilon^*() > 0$ s.t. the system is congested if and only if $\epsilon < \epsilon^*$

  - idea: ϵ^* is the mass of patients that would hold out for safe organs w/o death

Multiplicity and self-fulfilling congestion

Proposition

1. if there are multiple steady states, then they are Pareto ranked
2. all steady states except (possibly) the best one are congested

In fact, there can be congestion in one steady state and no congestion in a Pareto-dominating steady state ↝ “self-fulfilling congestion”

- intuition: say there are two types, healthy (low δ) and unhealthy (high δ)
- unhealthy waiting for safe can cause congestion w/o raising T_{safe} much
- this can raise T_{risky} enough (by more than change in T_{safe}) to get equilibrium

for remainder of the talk: focus on the best steady-state

Welfare effect of congestion

- congestion hurts patients who would accept risky kidneys by inflating T_{risky}
- in equilibrium, this makes more patients hold out for safe kidneys

Proposition

If the system is congested, then increasing ϵ strictly lowers both T_{safe} and T_{risky}

- so congestion (if it arises or worsens) makes *everyone* worse off

- all comparative statics via monotone comparative statics for equilibrium
  - Milgrom and Roberts (1994); also apply to the worst equilibrium

Congestion and constrained inefficiency

- how bad are the inefficiencies caused by congestion?
- does improving the supply of safe organs always improve welfare?

Proposition

Without organ expiry, increasing *s* strictly lowers T_{safe} and weakly lowers T_{risky}

Proposition

With organ expiry, there exist parameters for which increasing *s strictly raises* both T_{safe} and T_{risky}

- intuition: increasing *s* makes more people hold out for safe organs
- this worsens congestion, and can do so enough to raise waiting times
    - everyone worse off despite more people getting safe organs

Congestion and market thickness

- What happens if separate queuing systems are merged?
  - e.g., make “region” larger, or move from regional to national waitlists
- increases market thickness, but effects on congestion?

Proposition

If the system is congested, then increasing *p* strictly raises T_{safe} andT_{risky}

- intuition: if there is already congestion, having a thicker market worsens it
- and the system can transition from uncongested to congested if p increases
  - and will eventually if p increases enough

Delegating decisions to hospitals

- doctors may have different incentives than patients
  - penalty for adverse post-transplant outcomes (Schaefer et al. 22, Chan & Roth, 23)
  - Starting July 2023 pre-transplant mortalities penalties
  - Starting July 2024 acceptance rate penalties

Doctors' incentives and the congestion externality (I)

- suppose doctors of patients of type θ get utility

$$u_{doctor} = \Pr[\text{get safe organ}] + \lambda(\theta)Pr[\text{get risky organ}], \quad \text{where } \lambda(\theta) < \nu(\theta)$$

- so doctors' willingness to wait for a safe kidney is

$$W_\lambda(\theta) = \frac{-log\lambda(\theta)}{\delta(\theta)} > \frac{-log\nu(\theta)}{\delta(\theta)} = W_\nu(\theta)$$

Assume $W_\lambda(\theta)$ has an absolutely continuous distribution and is bounded

Doctors' incentives and the congestion externality (II)

Proposition

if the system is congested under patient decisions, then:

1. the steady state will also be congested under hospital decisions
2. if ν has full support on [0,1], then *all types of patients have strictly lower expected utility under hospital decisions* than under patient decisions
3. *total expected utility of hospitals is strictly lower* under hospital decisions than under patient decisions, with equality only if $P[\lambda(\theta) = 0] = 1$.

intuition: delegation to doctors makes more types hold out for safe organs,

which worsens congestion and increases waiting times

Congestion frictions - summary

- organ expiry can lead to congestion on deceased donor waiting lists
- causing substantial inefficiencies and externalities
- delegation to risk-averse doctors can worsen congestion and harm welfare
- possibility of self-fulfilling congestion that harms everyone

Potential remedies:

- expedite offers of low-quality organs, or create separate waiting lists by organ types (Castro et al., 2020)
- caution in expanding regions
- involve patients in rejection/acceptance decisions
- relax hospitals’ disincentives for accepting risky organs

Towards an adaptive policy

- Status quo:

  organs are allocation based on fixed priorities and fixed organs characteristics (e.g., KDPI)

- Information arrives over time
- Organ quality is revealed over time

Further informational and implementation challenges

- Information about an organ quality aggregates during time
    - Biopsy results, OR (clamp), imaging, refusal reasons from experts

- What makes an organ marginal and hard-to-place? How to place such organs?
- Can we identify quickly marginal organs? How to expedite the process for such organs? Need for an adaptive policy

*The following data from a working paper with Grace Guan, Mike Rees, Paulo Somaini and Alvin Roth

Increasing utilization of marginal organs

Refusals are signals but also part of communication

- Add information about quality and acceptance chances
- Refine/redesign information communication with centers
  - (e.g., ask if the center may accept it for any patient)

Expediting

- Classify marginal organs
- Batch offers
- Adapt KDRI during offering process
- Target aggressive centers

Summary

- organ expiry can lead to congestion on deceased donor waiting lists
- delegation to risk-averse doctors can worsen congestion and harm welfare
- possibility of self-fulfilling congestion that harms everyone

Possibilities:

- expedite offers of low-quality organs, or create separate waiting lists by organ types (Castro et al., 2020)
- Create separate lists by organ type (possibly with some information design)
- caution in designing local regions
- relax hospitals' disincentives for accepting risky organs