Greedy Matching of Impatient Agents:

The Role of Inventory

Angela Kohlenberg

Framework for understanding agent impatience in dynamic matching

Blood Transfusion

Ride-Sharing

Organ Transplantation

1. How costly is impatience? And...
2. What are the key levers for mitigating this cost?
    - **The Cost of Impatience in Dynamic Matching**, A. Kohlenberg and I. Gurvich, *Forthcoming at Management Science*
3. What matching policies minimize this cost?
    - **Greedy Matching of Impatient Agents**, A. Kohlenberg, *Under review at Management Science*

- Focus: Identify the key determinants of performance and optimal control

What matching policies minimize the match loss from abandonment (cost of impatience)?

Demand

Type O- (high priority/reward)

Others (low priority/reward)

Supply

Type O- (universal donor)

- **Match immediately:** Supply is not available for better future matches
- **High reward match only:** Supply may abandon
- **Optimal policy:** different in different markets $\Rightarrow$ key conditions

Key condition: It's all about inventory (of supply)

Demand

Type O- (high priority/reward)

Others (low priority/reward)

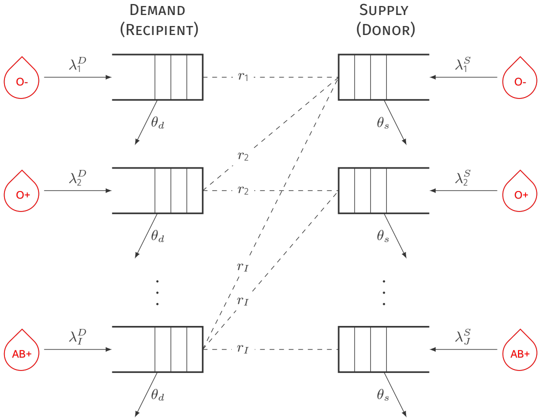
Supply

Type O- (universal donor)

Extreme cases:

1. Infinitely impatient supply
2. Infinitely impatient demand $\Rightarrow$ what if there is sufficient supply inventory?

What matching policies minimize the cost of impatience?



Agent impatience: demand impatience rate θ_d; supply impatience rate θ_s

Match reward: reward r_i obtained when demand type i matched with any supply type

Objective: Find matching policy, π, that maximizes long-run average reward, v^π

(Or, equivalently, minimizes reward loss due to impatience)

> What market conditions (parameters) favor waiting to match vs. matching immediately?

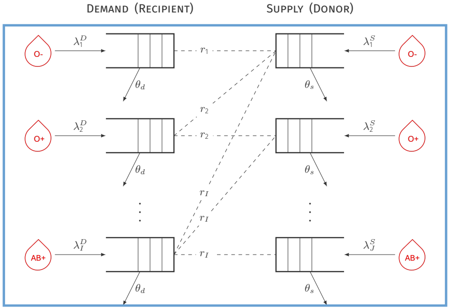
$\Rightarrow$ Optimality of (localized) greedy policies

Definition (Greedy Priority Policy, $G(\mathcal{P})$).

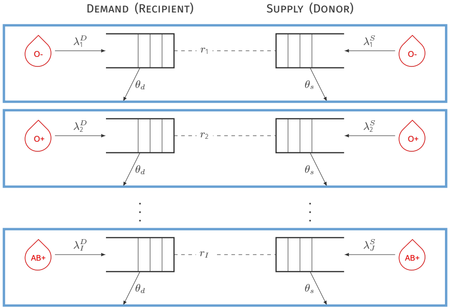
Perform an available match immediately on the network partition $\mathcal{P}$, with priority given to

- the **highest-reward** match in fully-connected networks, or
- the agent with the **longest queue** in partially-connected networks.

GREEDY MATCHING ON ENTIRE NETWORK, $G(\mathcal{I})$



HIGHEST-REWARD MATCH ONLY, $G(\mathcal{H}_1)$



Definition (Cost-of-Impatience under Policy π, CoI(π)).

$$\text{CoI}(\pi) := \text{Average no-abandonment reward } (\bar{v}) - \text{ Average reward under policy } \pi \ (v^{\pi})$$

Theorem (Scale Optimality of Policy $G(\mathcal{P})$).

Let $\mathcal{P}$ be the output of Algorithm 1. If $\mathcal{P} \neq \emptyset$, then for **"all" parameter combinations**, $G(\mathcal{P})$ is scale optimal:

$$\frac{\text{CoI}(G(\mathcal{P}))}{\text{CoI}(\pi^*)} = \frac{\bar{v} - v^{G(\mathcal{P})}}{\bar{v} - v^*} = \Theta(1) ,$$

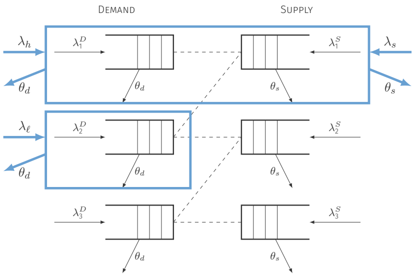
where $\pi^* = \arg\max_{\pi} v^{\pi}$ and $v^* = \max v^{\pi}$.

Further, for all network partitions $\overline{\mathcal{P}} \neq \mathcal{P}$, $G(\overline{\mathcal{P}})$ is not scale optimal.

- **Scale Optimality:** second-order measure of optimality
- **Constant Competitive Ratio:** $\frac{v^{\pi}}{v^*} = \frac{\bar{v} - \text{CoI}(G(\mathcal{P}))}{\bar{v} - \text{CoI}(\pi^*)} = \Theta(1)$

Algorithm to determine scale optimal network partition $\mathcal{P}$

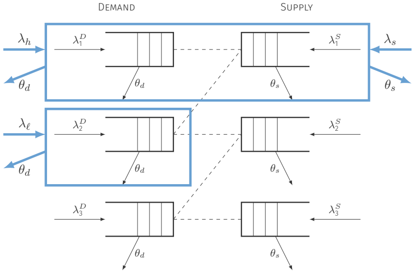
Iteration 1



Can and should supply-side inventory be held?

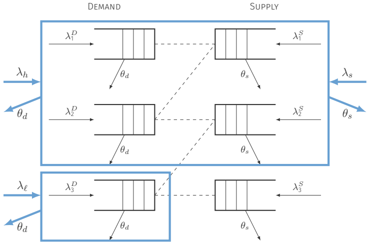
Algorithm to determine scale optimal network partition $\mathcal{P}$

Iteration 1



Supply
should never
be held

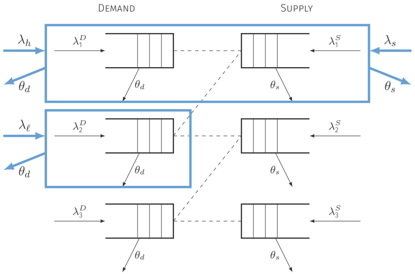
Iteration 2



Can and should
supply-side inventory
be held?

Algorithm to determine scale optimal network partition $\mathcal{P}$

Iteration 1

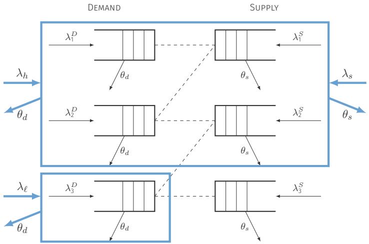


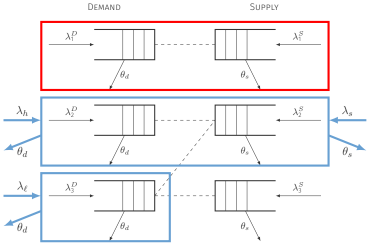
Can and should supply-side inventory be held?

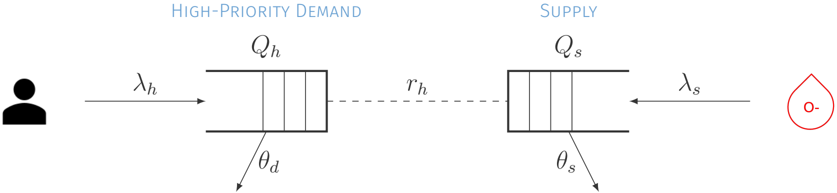
Supply **should never** be held

Supply **should always** be held

Iteration 2







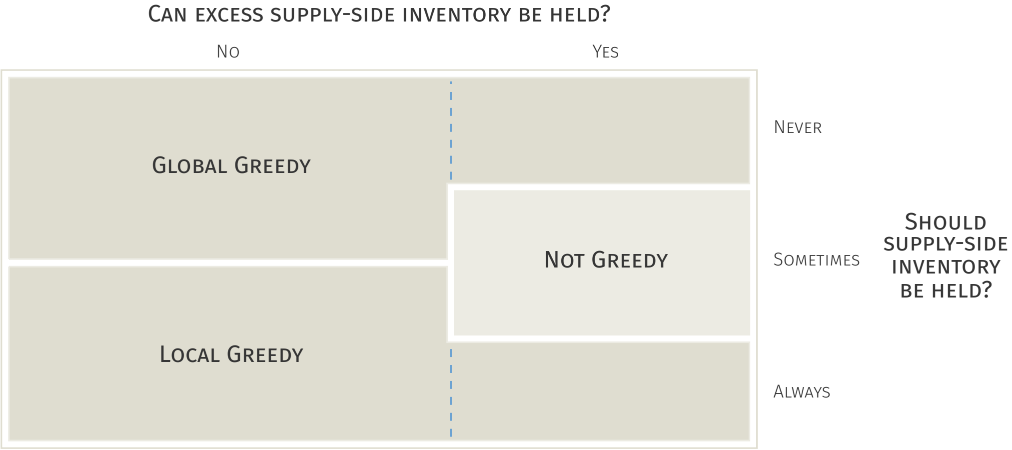
(Kohlenberg and Gurvich, 2024, Theorem 1, Universal Inventory Approximation).

For **"all" parameter combinations,**

$$\mathbb{E}[Q_h] = \Theta\left(\min\left\{\frac{\rho_h}{1-\rho_h}, \sqrt{\frac{\lambda_h}{\theta_d}}\right\}\left(1+\left[1+\frac{\rho_h}{1-\rho_h}\frac{\theta_d}{\lambda_h}\min\left\{\frac{\rho_h}{1-\rho_h}, \sqrt{\frac{\lambda_h}{\theta_d}}\right\}\right]\sqrt{\frac{\lambda_s}{\theta_s}}(1-\rho_h)e^{\frac{\lambda_s}{\theta_s}(1-\rho_h)^2\mathcal{F}(\rho_h)}\right)^{-1}\right)$$

where $\rho_h = \lambda_h/\lambda_s$ and $\mathcal{F}(\rho_h) = \sum_{n=1}^{\infty}\frac{1}{n(n+1)}(1-\rho_h)^{n-1}$.

Optimality of greedy policies: It's all about inventory



Can excess supply-side inventory be held?

Can excess supply-side inventory be held?

No	Yes
$\lambda_s - \lambda_h \leq \sqrt{\theta_s \lambda_s}$	$\lambda_s - \lambda_h \geq \sqrt{\theta_s \lambda_s}$

Can enough supply be held so that high-priority demand **always matches immediately** upon arrival?

- **Amount of Supply Inventory:** relative measure of **server patience** and **excess capacity**, $\lambda_s - \lambda_h$

Inventory cannot be held: Should supply-side inventory be held?

Can excess supply-side inventory be held?

No

High s/h-impatience ($\theta_s \geq \theta_d$)	
Low s/h-impatience ($\theta_s \leq \theta_d$)	

Does high-priority demand **abandon faster** than supply?

No

Should supply-side inventory be held?

Yes

- Inventory cannot be held: only relative patience matters

Inventory cannot be held: Greedy policy is always scale optimal

Can excess supply-side inventory be held?

No

Greedy (entire network) $G(\mathcal{I})$	
High-priority only $G(\mathcal{H}_1)$	

No (High s/h-impatience)

Should supply-side inventory be held?

Yes (Low s/h-impatience)

- **Inventory cannot be held:** $G(\mathcal{P})$ is always scale optimal, $\mathcal{P}$ depends only on mean patience

Inventory cannot be held: Greedy policy is always scale optimal

Can excess supply-side inventory be held?

No

Greedy (entire network)

$G(\mathcal{I})$

$\longleftarrow$

High-priority only

$G(\mathcal{H}_1)$

No (High s/h-impatience)

Should supply-side inventory be held?

Yes (Low s/h-impatience)

Inventory cannot and should not be held: high-priority demand will wait for supply & supply will not wait for high-priority demand

Inventory cannot be held: Greedy policy is always scale optimal

Can excess supply-side inventory be held?

No

Greedy (entire network)

$G(\mathcal{I})$

High-priority only

$G(\mathcal{H}_1)$ ⟵

No (High s/h-impatience)

Should supply-side inventory be held?

Yes (Low s/h-impatience)

Inventory cannot, but should be held: reserve some supply for high-priority demand; there is not enough supply to eventually match other demand types

Can excess supply-side inventory be held?

	Yes	
	High s/h-cost	Never
	Moderate s/h-cost	Sometimes
	Low s/h-cost	Always

Does high-priority demand have higher inventory **“holding cost”** than supply?

Should supply-side inventory be held?

Inventory holding cost: Combined measure of the **number of waiting agents**, their **patience rate**, and the **reward loss** for each agent abandonment

Inventory "Holding Cost"

Definition (Cost Ratio under Policy π, CR(π)).

$$\mathrm{CR}(\pi) := \frac{r_l\theta_s}{(r_h - r_l)\theta_d}\frac{\mathbb{E}^\pi[Q_s]}{\mathbb{E}^\pi[Q_h]} \in \left[\frac{r_l\lambda_s(1-\rho_h)}{(r_h - r_l)\theta_d \min\left\{\frac{\rho_h}{1-\rho_h}, \sqrt{\frac{\lambda_h}{\theta_d}}\right\}}, \frac{r_l\theta_s \min\left\{\frac{\rho_s}{1-\rho_s}, \sqrt{\frac{\lambda_s}{\theta_s}}\right\}}{(r_h - r_l)\theta_d \min\left\{\frac{\rho_h}{1-\rho_h}, \sqrt{\frac{\lambda_h}{\theta_d}}\right\}}\right]$$

"Cost" per unit held

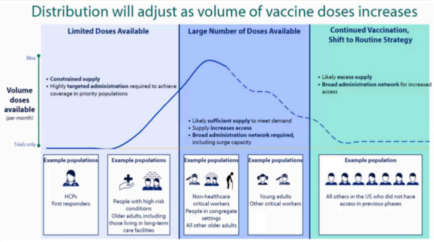
Number of units held

- Inventory under policy π: Approximation based on **Kohlenberg and Gurvich (2024)**
- High s/h-cost: lots of excess capacity, supply abandons much faster than demand, or very small difference in match rewards
- Low s/h-cost: very little excess capacity, demand abandons much faster than supply, or very large difference in match rewards

Inventory can be held: Greedy policy is scale optimal if one side has higher holding cost

Can excess supply-side inventory be held?

Yes

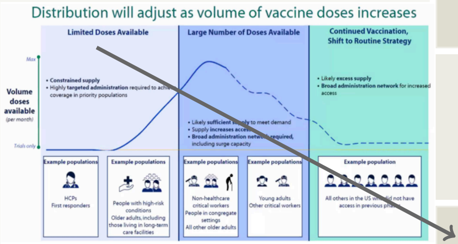


Greedy (entire network) $G(\mathcal{I})$	Never (High s/h-cost)	
No greedy policy	Sometimes	Should supply-side inventory be held?
High-priority only $G(\mathcal{H}_1)$	Always (Low s/h-cost)	

Inventory can be held: Greedy policy is scale optimal if one side has higher holding cost

Can excess supply-side inventory be held?

Yes



Greedy (entire network) $G(\mathcal{I})$ — Never (high s/h-cost)

No greedy policy — Sometimes

High-priority only $G(\mathcal{H}_1)$ — Always (low s/h-cost)

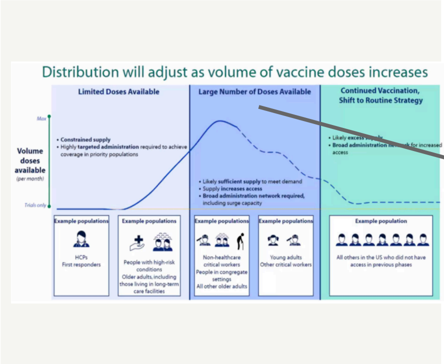
Should supply-side inventory be held?

Inventory can and should always be held: High cost to hold high-priority demand relative to supply; reserve all supply for high-priority demand

Inventory can be held: Greedy policy is scale optimal if one side has higher holding cost

Can excess supply-side inventory be held?

Yes

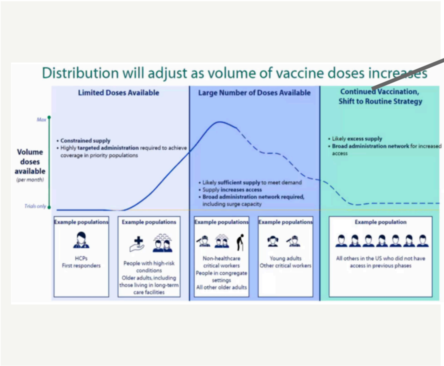


	Should supply-side inventory be held?
Greedy (entire network) $G(\mathcal{I})$	Never (high s/h-cost)
No greedy policy	Sometimes
High-priority only $G(\mathcal{H}_1)$	Always (low s/h-cost)

Inventory can and should sometimes be held: Reserve some supply for high-priority demand; perform lower-priority matches when there is enough supply inventory

Can excess supply-side inventory be held?

Yes



Greedy policy	Should supply-side inventory be held?
Greedy (entire network) $G(\mathcal{I})$	Never (high s/h-cost)
No greedy policy	Sometimes
High-priority only $G(\mathcal{H}_1)$	Always (low s/h-cost)

> **Inventory can but should never be held:** High cost to hold supply relative to high-priority demand; match supply immediately

Related literature (a sample): Dynamic matching with impatient agents

Optimality of Greedy Policies - Infinite Patience or Zero Patience

- **Greedy matching policy** is optimal in LARGE MARKETS (Ashlagi et al., 2023) and in suitably preprocessed two-way networks (Kerimov et al., 2023; Gupta, 2024)
- **Greedy policy strictly outperforms batch matching policy (waiting to match)** in some settings (Ashlagi et al., 2023)

No abandonment / Specific set of parameters

Optimal Algorithms - Random Abandonment

- **Batch matching policy** is optimal in HIGH-VOLUME SETTING: arrival rates scaled up and all other parameters held constant (Aveklouris et al., 2024)

Specific set of parameters

- **Batch matching can be bad** in some settings (Aouad and Saritaç, 2022)
- **Randomized matching policy** achieves constant approximation of upper bound on AVERAGE REWARD (Aouad and Saritaç, 2022; Collina et al., 2020)

Single policy for all sets of parameters

▌ Our Focus: universal results (in parameters) & general characterization

▌ Thank you!

Extra Slides

References

Aouad, A. and Saritaç, Ö. (2022). Dynamic stochastic matching under limited time. *Operations Research*, 70(4):2349–2383.

Ashlagi, I., Nikzad, A., and Strack, P. (2023). Matching in dynamic imbalanced markets. *The Review of Economic Studies*, 90(3):1084–1124.

Aveklouris, A., DeValve, L., Stock, M., and Ward, A. (2024). Matching impatient and heterogeneous demand and supply. *Operations Research*, page forthcoming.

Collina, N., Immorlica, N., Leyton-Brown, K., Lucier, B., and Newman, N. (2020). Dynamic weighted matching with heterogeneous arrival and departure rates. In *Web and Internet Economics: 16th International Conference, WINE 2020, Beijing, China, December 7–11, 2020, Proceedings 16*, pages 17–30. Springer.

Gupta, V. (2024). Greedy algorithm for multiway matching with bounded regret. *Operations Research*, 72(3):1139–1155.

Kerimov, S., Ashlagi, I., and Gurvich, I. (2023). On the optimality of greedy policies in dynamic matching. *Operations Research*, page forthcoming.

Kohlenberg, A. and Gurvich, I. (2024). The cost of impatience in dynamic matching: Scaling laws and operating regimes. *Management Science*, page forthcoming.

Summary: What matching policies minimize the cost of impatience?