A Mechanism Design Approach to "Gainers and Losers in Priority Services"*

Peiran Xiao†

July 2024

Abstract

I reformulate Gershkov and Winter's (2023) model of priority services as a mechanism design problem under a feasibility condition. Thus, I provide the necessary and sufficient conditions for their Propositions 1, 2, 7, and 8, while allowing for stochastic priority levels. Under the weaker conditions, adding more priority levels can increase both the provider's revenue and consumer welfare if the cost distribution has a decreasing failure rate but satisfies Myerson's regularity. Full separation can be implemented by an all-pay auction. I also show that the provider can guarantee at least half the maximal revenue by offering one priority service in addition to a (free) regular service, and the approximation can be arbitrarily close if the distribution is sufficiently concave. The approach can also be applied to Moldovanu et al.'s (2007) model of status contests.

Gershkov and Winter (2023, henceforth GW) study a model of priority service (PS) and analyze its welfare implications on consumers. They show that introducing PS can be detrimental to consumer welfare when the monopoly service provider extracts more revenue than the efficiency gains. For multiple priority levels, they find that if the distribution has an increasing failure rate (IFR), the provider's revenue is strictly increasing in the number of priority levels, while the customers' welfare is higher if PS is not offered than if multiple priority levels are offered.

In this comment, I reformulate their model (with a monopoly service provider) as a mechanism design problem under a feasibility condition. Thus, I rewrite the revenue and consumer welfare maximization problems as linear optimization problems under

*I thank Alex Gershkov, Bart Lipman, and Juan Ortner for their helpful comments.
†Department of Economics, Boston University. Email: pxiao@bu.edu.

majorization constraints. Using the techniques in Kleiner et al. (2021, henceforth KMS), I provide the necessary and sufficient conditions for GW's Propositions 1, 2, 7, and 8, while allowing for stochastic priority levels. In particular, consumer welfare is decreasing (increasing) in the number of priority levels if the failure rate is increasing (decreasing), while an increasing failure rate is sufficient but unnecessary for the provider's revenue to be increasing in the number of priority levels. Thus, the trade-off between the provider's revenue and consumer welfare maximization is less stark—if the distribution has a decreasing failure rate but satisfies Myerson's regularity condition (e.g., Pareto distribution), adding more priority levels can increase both. Full separation can be implemented by an all-pay auction. Moreover, I also show that the provider can guarantee at least half the maximal revenue by offering two priority levels—a priority service and a (free) regular service. The approximation can be arbitrarily close if the distribution is sufficiently concave.

In addition, I extend their Section IV which considers endogenous pricing of the regular service (and hence exclusion) to multiple priority levels. Under some regularity conditions, the revenue-maximizing (welfare-maximizing) mechanism excludes some types and induces full separation (total pooling) of the rest.

This approach can also be applied to other allocation problems where allocating the object to one agent has externalities on others, for example, the model of contests for status by Moldovanu et al. (2007, henceforth MSS) (with a continuum of agents). Among others, I obtain the necessary and sufficient condition for their Theorem 4.

1 Model

In their benchmark model, a monopoly service provider faces a continuum of customers of mass 1 with heterogeneous unit costs of waiting (i.e., types), denoted by c, which has a distribution F with bounded support $[0, \bar{c}]$ and density $f(c) > 0$. A customer with type c who gets service at time t and pays $p \geq 0$ has a utility of $-p - tc$. The provider can serve a single customer at each instant. They normalize the service time of a mass m of consumers to be exactly m units of time and the cost of the provider to be zero, which implies $t \in [0, 1]$ because the total mass of customers is 1.

They assume the regular service is free and do not consider exclusion in the benchmark model. Then, in Section IV, they allow for endogenous pricing of the regular service (and thus exclusion) by introducing the value of the service to the consumer. I maintain this assumption first and consider endogenous pricing of the regular service and exclusion in Section 4, as in GW's Section IV.

2 A Mechanism Design Approach

It is without loss to consider a direct mechanism $\{p(c), t(c)\}$, consisting of a payment $p\colon [0,\bar{c}] \to \mathbb{R}_+$ and a (potentially random) ex-post waiting time $T\colon [0,\bar{c}] \times \Omega \to [0,1]$, where $\omega \in \Omega$ captures the randomness, and $t(c) = \mathbf{E}[T(c,\omega) \mid c]$. Denote $U(c) = -p(c) - t(c)c$.

Lemma 1. *A direct mechanism $\{p(c), t(c)\}$ is incentive-compatible if and only if*

- $U(c) = U(0) - \int_0^c t(x)\,\mathrm{d}x$ *for all* $c \in [0,\bar{c}]$*, and*
- $t(c)$ *is decreasing.*

Alternative interpretation. The provider designs a queuing scheme $\ell\colon [0,\bar{c}] \to \Delta L$ that maps the consumer's type c to a (possibly stochastic) priority level $\ell(c)$ along with a payment $p\colon [0,\bar{c}] \to \mathbb{R}_+$; the higher the priority level, the less time he needs to wait.[1] If there are $k \geq 1$ priority levels, $L = \{1,\ldots,k\}$. For full separation, $L = [0,1]$. When priority levels are *deterministic*, $\ell(c) \in L$ is degenerate. By the assumptions on the service time, for a given level $\ell_0 \in L$, the expected waiting time conditional on ℓ_0 is $\tau(\ell_0) = \Pr(\ell(c) > \ell_0) + \Pr(\ell(c) = \ell_0)/2$. For deterministic priority levels, incentive compatibility requires that $\ell(c)$ must be (weakly) increasing, so the consumer's expected waiting time is $t(c) = \tau(\ell(c)) = 1 - \mathbf{E}[F(\tilde{c}) \mid \ell(\tilde{c}) = \ell(c)]$.[2] For potentially stochastic priority levels, the expected waiting time is $t(c) = \mathbf{E}_{\tilde{\ell}\sim\ell(c)}[\tau(\tilde{\ell})] = 1 - \mathbf{E}_{\tilde{\ell}\sim\ell(c)}[\mathbf{E}[F(\tilde{c})|\tilde{\ell}]]$.[3]

A decreasing waiting time function $t(c)$, albeit incentive-compatible, is not necessarily feasible because it may be unable to be induced by an ex-post allocation (or queuing scheme). Formally, say $t(c)$ is *feasible* if there exists an ex-post allocation T (or a queuing scheme ℓ) that induces $t(c)$ as the interim waiting time. Now I characterize the necessary and sufficient condition for a decreasing $t(c)$ to be feasible.

It is convenient to denote the consumer's priority value by $s(c) = 1 - t(c)$, which quantifies priority by the time he saved compared to being served last. In other words, $s(c) = \mathbf{E}_{\tilde{\ell}\sim\ell(c)}[\mathbf{E}[F(\tilde{c})|\tilde{\ell}]]$, where $\ell(c)$ is his (relative) priority *level.*

Theorem 1 (Feasibilty)**.** *A decreasing $t(c)$ is feasible if and only if $s(c) = 1 - t(c)$ is a*

[1] This can be implemented by a pricing scheme as a function of the priority level.

[2] To see this, if $\ell(c)$ is strictly increasing at c, then $t(c) = 1 - F(c)$. Otherwise, if $\ell(c) = \ell_1$ is constant if and only if $c \in [c_1, c_2]$ (i.e., $[c_1, c_2]$ are assigned to the same level ℓ_1), then $t(c) = 1 - (F(c_1) + F(c_2))/2 = 1 - \mathbf{E}[F(\tilde{c}) \mid \tilde{c} \in [c_1, c_2]] = 1 - \mathbf{E}[F(\tilde{c}) \mid \ell(\tilde{c}) = \ell(c)]$ for $c \in [c_1, c_2]$.

[3] For deterministic priority levels, it can be viewed as a monotone categorization problem (see Rayo (2013) and Onuchic and Ray (2023)). See also Xiao (2024) for potentially stochastic levels.

mean-preserving spread of $F(c)$ in the quantile space, denoted by $s \in \mathrm{MPS}(F)$, that is,

$$\int_x^{\bar{c}} s(c)\,\mathrm{d}F(c) \leq \int_x^{\bar{c}} F(c)\,\mathrm{d}F(c) \textit{ for all } x \in [0, \bar{c}], \tag{1}$$

$$\int_0^{\bar{c}} s(c)\,\mathrm{d}F(c) = \int_0^{\bar{c}} F(c)\,\mathrm{d}F(c) = 1/2. \tag{2}$$

Proof sketch. À la Kleiner et al. (2021, Theorem 3). □

Remark 1. The condition resembles the symmetric version of Border's Theorem in reduced-form auctions (see Maskin and Riley, 1984; Matthews, 1984; Border, 1991).

Remark 2. If the priority levels $\ell(c)$ are deterministic, $s(c) = 1 - t(c)$ can only be an extreme point of $\mathrm{MPS}(F)$, i.e., either the majorization constraint or the monotonicity constraint binds (see Theorem 1 in KMS). Since I allow for stochastic priority levels, $t(c)$ can take other forms.

Remark 3. An analog of the theorem also applies to the model of contests for status by MSS with a continuum of agents, in which a type-θ agent assigned to the status level $\ell(\theta)$ has a status value of $s(\theta) = \mathbf{E}_{\tilde{\ell}\sim\ell(\theta)}[\mathbf{E}[F(\tilde{\theta})|\tilde{\ell}] - \mathbf{E}[1 - F(\tilde{\theta})|\tilde{\ell}]]$, that is, the number of agents behind him subtracted by the number of agents ahead of him.[4] Thus, the feasibility condition on the value of status becomes: an increasing $s(\theta)$ is feasible if and only if $s \in \mathrm{MPS}(2F - 1)$ in the quantile space.[5] Because $\mathbf{E}[s] = \mathbf{E}[2F - 1] = 0$, the total value is normalized to 0.

The intuition is that the provider can induce full separation (i.e., $t(c) = 1 - F(c)$) by offering infinitely many priority levels and serving every type in the descending order of their costs. Any pooling of different types into the same priority level makes $s \equiv 1 - t$ a mean-preserving spread of F in the quantile space.

Example. Suppose there are two priority levels—regular service ($\{p_L, t_L\}$) and priority service ($\{p_H, t_H\}$), where $p_L = 0$ (no exclusion). Denote the cutoff type by c^*. Then, the waiting time is $t_L = 1 - F(c^*)/2$ for $c < c^*$ and $t_H = (1 - F(c^*))/2$ for $c \geq c^*$. It is easy to

[4]To see this, the expected status value conditional on the status level being ℓ_0 is $\Pr(\ell(\theta) > \ell_0) - \Pr(\ell(\theta) < \ell_0)$, and incentive compatibility requires that $\ell(\theta)$ is decreasing (because the marginal cost is $1/\theta$).

[5]MSS consider a finite number of n agents. In that case, under full separation (n status levels), agent i's status value $s(\theta_i)$ is

$$\frac{1}{n}\sum_{k=1}^{n}\binom{n}{k}(1 - F(\theta_i))^{n-k}F(\theta_i)^k[k - (n - k - 1)] = 2F(\theta_i) - \frac{n+1}{n} \to 2F(\theta_i) - 1 \tag{3}$$

as $n \to \infty$.

check that $\mathbf{E}[t] = 1/2$ and that $\int_0^c t(c')\,\mathrm{d}F(c') \le \int_0^c (1 - F(c'))\,\mathrm{d}F(c')$ for all $c \in [0, \bar{c}]$ (with equality at c^* and $\bar{c}$).

3 Optimal Mechanism

Two extreme mechanisms are particularly of interest: full separation ($t(c) = 1 - F(c)$) and total pooling ($t(c) = 1/2$)[6] ; the former can be induced by offering as many priority service levels as possible, and the latter can be induced by offering no priority service (regular service only). Note that I allow for stochastic priority levels.[7]

For this section, I maintain GW's assumption in the benchmark model that the regular service $\{p_L, t_L\}$ is free ($p_L = 0$) so that there is no exclusion. Thus, the type-c consumer's reservation utility is $\underline{U}(c) = -t_L c \le -tc$, where t_L is the expected waiting time of the regular service, which is endogenously determined by the scheme $t(c)$—i.e., $t_L = \max_{c \in [0,\bar{c}]} t(c)$.

3.1 Revenue maximization

I first consider the optimal mechanism that maximizes the provider's revenue. The revenue maximization problem is given by

$$\max_{t(c),p(c)} \int_0^{\bar{c}} p(c)\,\mathrm{d}F(c) \tag{4}$$

subject to

$$p(c) \ge 0 \tag{5}$$

$$U(c) \equiv -p(c) - t(c)c \ge -t_L c \qquad \text{(IR)} \tag{6}$$

$$-p(c) - t(c)c = U(0) - \int_0^c t(x)\,\mathrm{d}x \qquad \text{(IC)} \tag{7}$$

$$t(c) \text{ is decreasing} \tag{8}$$

$$s \equiv 1 - t \in \mathrm{MPS}(F) \qquad \text{(MPS)} \tag{9}$$

Note that (IR) can be written as $\tilde{U}(c) \equiv U(c) + t_L c \ge 0$, which satisfies $\tilde{U}'(c) = t_L - t(c) \ge 0$, so (IR) reduces to $U(0) \ge 0$. Further, because $p'(c) = -t'(c)c \ge 0$, the constraint $p(c) \ge 0$

[6] Both are the extreme points of $\mathrm{MPS}(F)$ (in the quantile space).

[7] GW also assume (except in Section V) that the provider can offer at most two priority levels—regular service ($\{p_L, t_L\}$) and priority service ($\{p_H, t_H\}$) where $p_L = 0$. While I allow for multiple priority levels in the first place, the results also apply under the restriction to two priority levels.

reduces to $p(0) \geq 0$.

The revenue is given by

$$R = \int_0^{\bar{c}} p(c)\,\mathrm{d}F(c) = \int_0^{\bar{c}} \left(\frac{1-F(c)}{f(c)} - c\right) t(c)\,\mathrm{d}F(c) - U(0). \tag{10}$$

Because $U(0) = -p(0) \geq 0$ (IR), it is optimal to set $U(0) = p(0) = 0$ (even if p were allowed to be negative). Denote $J(c) = c - \frac{1-F(c)}{f(c)}$. The revenue maximization problem is equivalent to

$$\max_{s \in \mathrm{MPS}(F)} \int_0^{\bar{c}} J(c)s(c)\,\mathrm{d}F(c) \tag{11}$$

where $s(c) = 1 - t(c)$ is increasing.

Note that I maintain the assumption that the regular service is free so that there is no exclusion. Proposition 2 in KMS immediately implies the following results.

Proposition 1 (Cf. GW's Proposition 8)**.** *Adding more priority levels (strictly) increases the provider's revenue if and only if $J(c) = c - \frac{1-F(c)}{f(c)}$ is (strictly) increasing.*

Remark 4. GW's Proposition 8 provides a sufficient condition for this result: F satisfies the IFR property (i.e., $\frac{1-F(c)}{f(c)}$ is decreasing).

Corollary 1.1 (Cf. MSS's Theorem 4)**.** *The revenue-maximizing mechanism induces full separation if and only if $J(c) = c - \frac{1-F(c)}{f(c)}$ is increasing, which is payoff equivalent to an all-pay auction.*

Remark 5. Because MSS assume linear effort costs, effort maximization in their model is the same as revenue maximization here. In Theorem 4, they provide a sufficient condition for full separation: F satisfies the IFR property.

Full separation can be implemented by an all-pay auction, in which the more money a consumer pays, the more ahead he is in the line (and the less time he needs to wait).

Proposition 2 (Cf. MSS's Theorem 3)**.** *The revenue-maximizing mechanism always separates the highest types $c \in [\bar{c} - \varepsilon, \bar{c}]$ for some $\varepsilon > 0$.*

Proof sketch. Because $J'(\bar{c}) = 2 > 0$, $J(c)$ is increasing at a neighborhood of $\bar{c}$, so by Proposition 2 in KMS, the revenue-maximizing mechanism always separates types in this neighborhood. □

Remark 6. This "separation at the top" result is a continuous-type analog of MSS's Theorem 3, which assumes finite agents.

Proposition 3 (Cf. MSS's Proposition 1)**.** *If $J(c)$ is single-dipped (i.e., $\int_0^c J(x)\,\mathrm{d}x$ is concave-convex), there exists some $c_0 \in [0, \bar{c})$ such that the revenue-maximizing mechanism separates high types $c \in [c_0, \bar{c}]$ and pools types $c \in [0, c_0]$.*

If $F(c)$ is sufficiently concave, $J(c)$ is single-dipped, and the separating region $[c_0, \bar{c}]$ can be arbitrarily small.

Proof sketch. The first part follows from KMS. The second part follows from $J'(c) = 2 + (1 - F(c))f'(c)/f(c)^2$ and that a more concave F has a smaller (more negative) $f'/f < 0$. □

The proposition implies that offering two priority levels (i.e., regular and priority services) can perform arbitrarily closely to the revenue-maximizing mechanism if $F(c)$ is sufficiently concave.

Remark 7. The second half is a continuous-type analog of MSS's Proposition 1.

Remark 8. In the finite-agent version of GW,[8] the proposition implies that offering two priority levels (i.e., priority and regular services) maximizes the provider's revenue if $F(c)$ is sufficiently concave. PS is only sold to the agent of the highest type (VIP).

Proposition 4. *The provider can obtain at least half the maximal revenue by offering two priority levels (i.e., priority and regular services).*

Proof. By GW, when the price of PS is p, the cutoff type indifferent between both levels is

$$-p - c^*(p)\frac{1 - F(c^*(p))}{2} = -c^*(p)\left(1 - \frac{F(c^*(p))}{2}\right) \iff c^*(p) = 2p.$$

Thus, the maximal revenue by offering two levels is $\max R_2 \equiv \max_p p(1 - F(2p)) = \max_p p(1 - F(p))/2$. Consider the auxiliary screening problem of selling an indivisible item to one buyer, in which a standard result implies a simple posted-price mechanism is optimal (see Börgers, 2015, Proposition 2.5). In other words, denote $\mathcal{M} = \{q \colon [0, \bar{c}] \to [0, 1] \mid q \text{ increasing}\}$, then $\max_{q \in \mathcal{M}} \int_0^{\bar{c}} J(c)q(c)\,\mathrm{d}F(c) = \max_p p(1 - F(p))$, and any maximizer q^* must be an extreme point of $\mathcal{M}$ (i.e., a posted-price mechanism). Because $\mathrm{MPS}(F) \subseteq \mathcal{M}$, we have

$$\max R = \max_{s \in \mathrm{MPS}(F)} \int_0^{\bar{c}} J(c)s(c)\,\mathrm{d}F(c) < \max_{q \in \mathcal{M}} \int_0^{\bar{c}} J(c)q(c)\,\mathrm{d}F(c) = \max_p p(1 - F(p)) = 2 \max R_2.$$

[8] The provider will need to learn the set of agents' type realizations before setting the priority levels and payment schedule. This assumption is not required in MSS because the principal does not need to design the payment (instead, agents "pay" their effort costs autonomously).

The inequality is strict because an extreme point of $\mathcal{M}$ (i.e., a posted-price mechanism) cannot be a mean-preserving spread of F in the quantile space. $\square$

Remark 9. The result on the lower bound does not require the IFR property or the regularity condition.

Intuitively, in the auxiliary problem of selling a (continuum of) indivisible item to a (continuum of) consumer, selling an item to one consumer has no externalities on others. By contrast, negative externalities are present when selling priority levels (or allocating status)—for example, the seller cannot offer the highest priority to all at face value. Thus, the latter is subject to an additional constraint, resulting in a (strictly) less revenue.

3.2 Consumer welfare maximization

The consumer welfare maximization problem is given by

$$\max_{t(c),p(c)} \int_0^{\bar{c}} U(c)\,\mathrm{d}F(c) \tag{12}$$

subject to constraints (5)–(9). The customer's welfare can be written as

$$W = \int_0^{\bar{c}} U(c)\,\mathrm{d}F(c) = \int_0^{\bar{c}} \left(-\frac{1-F(c)}{f(c)}\right) t(c)\,\mathrm{d}F(c) + U(0). \tag{13}$$

Because $p(0) \geq 0$, $U(0) = -p(0) \geq 0$ (IR) implies $U(0) = p(0) = 0$.[9] The revenue maximization problem is equivalent to

$$\max_{s\in\mathrm{MPS}(F)} \int_0^{\bar{c}} \left(\frac{1-F(c)}{f(c)}\right) s(c)\,\mathrm{d}F(c) - \mathbf{E}[c] \tag{14}$$

where $s(c) = 1-t(c)$ is increasing. Proposition 2 in KMS immediately implies the following results.

Proposition 5 (Cf. GW's Proposition 7)**.** *Adding more priority levels increases the customers' welfare if and only if* $\frac{1-F(c)}{f(c)}$ *is increasing and decreases the customers' welfare if* $\frac{1-F(c)}{f(c)}$ *is decreasing. Without priority service, the customers' welfare is* $-\mathbf{E}[c]/2$.

Remark 10. GW's Proposition 7 shows the sufficiency of IFR property ($\frac{1-F(c)}{f(c)}$ is decreasing) for no priority service to be consumer welfare-maximizing. Proposition 5 shows

[9] This would not be true without the assumption $p(c) \geq 0$. If negative transfers are allowed and subject to budget balance, full separation (assortative matching) will maximize consumer welfare because it is efficient, and transfers can redistribute the efficiency gains among consumers (see, e.g., Gershkov and Schweinzer, 2010). I thank Alex Gershkov for pointing this out.

under IFR, adding more priority levels decreases consumer welfare, so offering no PS maximizes consumer welfare. Proposition 6 will provide a *necessary* and sufficient condition.

Corollary 5.1. *The consumer welfare-maximizing mechanism*

(i) offers no priority service (i.e, total pooling) if $\frac{1-F(c)}{f(c)}$ *is decreasing (IFR);*

(ii) induces full separation if and only if $\frac{1-F(c)}{f(c)}$ *is increasing.*

Remark 11. Part (ii) provides the necessary and sufficient condition for the counterpart of GW's Proposition 7—i.e., the optimality of full separation.

Corollary 5.2. *If both* $c - \frac{1-F(c)}{f(c)}$ *and* $\frac{1-F(c)}{f(c)}$ *are increasing (e.g., Pareto and Weibull distribution), adding more priority levels can increase both the provider's revenue and consumer welfare.*[10] *Thus, full separation is both revenue- and welfare-maximizing.*

Proposition 6 (Cf. GW's Proposition 1 and 7)**.** *Offering no PS (i.e., total pooling) maximizes the consumer welfare if and only if* $\int_0^c \left(\frac{1-F(x)}{f(x)} - \mathbf{E}[c]\right) \mathrm{d}F(x) \geq 0$ *for all* $c \in [0, \bar{c}]$.

Proof. Denote $H(c) = \int_0^c \frac{1-F(x)}{f(x)} \mathrm{d}F(x)$. Then, the condition in the proposition is equivalent to $H(c) \geq H(\bar{c})F(c) = \mathbf{E}[c]F(c)$ (graphically, $H(c)$ lies above the line connecting $H(0) = 0$ and $H(\bar{c}) = \mathbf{E}[c]$ in the quantile space). Therefore, $H(\bar{c})F(c)$, which corresponds to total pooling, is the convex hull of H (i.e., the largest convex function that lies below H).[11] By Proposition 2 in KMS, this condition is necessary and sufficient for total pooling (i.e., when PS is not offered) to be welfare-maximizing. □

Remark 12. Proposition 6 provides a necessary and sufficient condition for GW's Proposition 7. Under this condition, customers' welfare is higher when PS is not offered than when *any* $k > 1$ levels of PS are offered.

GW's Proposition 1 provides a stronger sufficient condition: $\mathbf{E}[c] - \frac{1-F(c)}{f(c)}$ "changes sign at most once, from negative to positive" (i.e., single-crosses zero from below) for consumer welfare to be higher when PS is not offered than when *one level of* PS is offered, which implied by the result above. Because $\int_0^{\bar{c}} \frac{1-F(c)}{f(c)} \mathrm{d}F(c) = \mathbf{E}[c]$, their single-crossing condition (akin to the first-order stochastic dominance) implies the condition in above (akin to the second-order stochastic dominance).

[10] The Pareto distribution has a CDF $F(x) = 1 - (\beta/x)^\alpha$. When $\alpha > 1$ and $\beta > 0$, both $\frac{1-F(x)}{f(x)}$ and $x - \frac{1-F(x)}{f(x)}$ are strictly increasing. The Weibull distribution has a CDF $G(x) = 1 - \exp(-x^k)$. When $k < 1$, both $\frac{1-F(x)}{f(x)}$ and $x - \frac{1-F(x)}{f(x)}$ are strictly increasing.

[11] A piecewise affine function consisting of other points on $H(c)$, which corresponds to offering more priority levels, lies above $H(\bar{c})F(c)$ and leads to lower consumer welfare.

Remark 13. A sufficient condition is the IFR property, according to Proposition 5.

Moreover, for the highest type $\bar{c}$, maximizing his utility $U(\bar{c})$ subject to constraints (5)–(9) is equivalent to

$$\max_{s \in \mathrm{MPS}(F)} U(0) + \int_0^{\bar{c}} \frac{s(c)-1}{f(c)} \, \mathrm{d}F(c). \tag{15}$$

Because $p(0) \geq 0$, (IR) implies $U(0) = p(0) = 0$. By Proposition 2 in KMS, if f is increasing, $s(c) = 1/2$ (i.e., total pooling) maximizes $U(\bar{c})$. Under total pooling, $U(c)$ is linear (because $U'(c) = -1/2$), and thus $U(c) = U(\bar{c}) \cdot c/\bar{c}$ for all $c \in [0, \bar{c}]$. With more priority levels, $U(c)$ is convex (while $U(0) = 0$ still holds), so $U(c) \leq U(\bar{c}) \cdot c/\bar{c}$ for all $c \in [0, \bar{c}]$. Therefore, $s(c) = 1/2$ also maximizes $U(c)$ for all $c \in [0, \bar{c}]$.

Proposition 7 (Cf. GW's Proposition 2)**.** *Offering no priority service (total pooling) maximizes every consumer's utility if and only if $F(c) \leq c/\bar{c}$ (i.e., F first-order stochastic dominates the uniform distribution). A sufficient condition is that $f(c)$ is increasing.*

Proof sketch. $U(\bar{c}) = \int_0^{\bar{c}} \frac{s(c)-1}{f(c)} \, \mathrm{d}F(c)$. Note that $\int_0^c 1/f(x) \, \mathrm{d}F(x) = c$, and the condition is equivalent to $\int_0^c (1/f(x) - \bar{c}) \, \mathrm{d}F(x) \geq 0$. The rest is similar to the proof of Proposition 6. □

Remark 14. GW's Proposition 2 shows that if $F(c) \leq c/\bar{c}$, all consumers are worse off after the introduction of one level of PS ($k = 2$) than $k = 1$. Proposition 7 shows this condition is necessary and sufficient for all consumers to be worse off after *any* $k > 1$ levels of PS are offered than if no PS is offered ($k = 1$).

3.3 Discussion on Exclusion

To this point, I have maintained the assumption that the regular service is free so that there is no exclusion.[12] If exclusions are possible, it will affect the feasibility of the waiting time function $t(c)$ by requiring $s(c) = 1 - t(c)$ to be a mean-preserving spread of $F(c)$ (in the quantile space) for the included types only. For example, if the provider excludes $c \in [0, \hat{c}]$, then $s = 1 - t \in \mathrm{MPS}(F)$ on $[\hat{c}, \bar{c}]$ (in the quantile space).[13] This effect on feasibility is also present in the standard monopoly problem (e.g. Mussa and Rosen,

[12]MSS do not consider exclusion either, while adding more status levels also lowers the utility of the lower types (who have higher costs).

[13]In other words, if low types are excluded, the feasibility condition in Theorem 1 becomes: s is *weakly* majorized by F, denoted by $s \in \mathrm{MPS}_w(F)$, that is,

$$\int_x^1 s(c) \, \mathrm{d}F(c) \geq \int_x^1 F(c) \, \mathrm{d}F(c),$$

and the equality need not hold at $x = 0$.

1978), where the revenue-maximizing seller will exclude agents with negative marginal revenue ($J(\theta) < 0$), while no agents will be excluded in welfare maximization. However, the participation constraint is trickier here because the consumer's reservation utility, which equals the value of the service to him, is ambiguous. This calls for introducing the value of the service to the model, as in GW's Section IV.

4 Endogenous Pricing of the Regular Service

In GW's Section IV, they allow for endogenous pricing of the regular service (and hence exclusion) and introduce an additional variable—the value of the service to the consumer. They assume the consumer has a unidimensional type $\theta \in [0, \bar{\theta}]$ that determines their unit cost of waiting, $c(\theta)$, and value of the service, $v(\theta)$. The consumer's utility is $v(\theta) - p - tc(\theta)$. Abusing notations, denote the CDF of type θ by $F(\theta)$ and the density by $f(\theta) > 0$ on the support $[0, \bar{\theta}]$. Denote $U(\theta) = v(\theta) - p(\theta) - t(\theta)c(\theta)$.

I maintain their assumptions that $v(\theta) > 0$, $c(\theta) > 0$, $v'(\theta) \geq 0$, $c'(\theta) \geq 0$, $v''(\theta) \leq 0$, $c''(\theta) \geq 0$, and $v'(\theta) > c'(\theta)$ (for low-type exclusion in their Section IV.A).[14] I further assume $v(0) - c(0) \geq 0$.

Lemma 2. *A direct mechanism $\{p(\theta), t(\theta)\}$ is incentive-compatible if and only if*

- *there exists a cutoff type $\hat{\theta} \in [0, \bar{\theta}]$ such that $U(\theta) \geq 0$ if and only if $\theta \geq \hat{\theta}$,*
- $U(\theta) = U(\hat{\theta}) + \int_{\hat{\theta}}^{\theta} (v'(x) - t(x)c'(x))\,\mathrm{d}x$ *for all $\theta \in [\hat{\theta}, \bar{\theta}]$, and*
- $t(\theta)$ *is decreasing.*

Note that (IC) implies $U'(\theta) = v'(\theta) - t(\theta)c'(\theta) > 0$ because $v'(\theta) > c'(\theta)$ and $t(\theta) \in [0, 1]$, so only low types will be excluded.

4.1 Revenue maximization

The revenue maximization problem is given by

$$\max_{t(\theta), p(\theta), \hat{\theta}} \int_{\hat{\theta}}^{\bar{\theta}} p(\theta)\,\mathrm{d}F(\theta) \tag{16}$$

[14] In Section IV.B, they assume $v'(\theta) < c'(\theta)$, which leads to high-type exclusion.

subject to the following constraints on $[\hat{\theta}, \bar{\theta}]$:

$$p(\theta) \geq 0 \tag{17}$$

$$U(\theta) \equiv v(\theta) - p(\theta) - t(\theta)c(\theta) \geq 0 \qquad \text{(PC)} \tag{18}$$

$$v(\theta) - p(\theta) - t(\theta)c(\theta) = U(\hat{\theta}) + \int_{\hat{\theta}}^{\theta} (v'(x) - t(x)c'(x))\,\mathrm{d}x \qquad \text{(IC)} \tag{19}$$

$$t(\theta) \text{ is decreasing} \tag{20}$$

$$s \equiv 1 - t \in \mathrm{MPS}_w(F) \qquad \text{(MPSw)} \tag{21}$$

Denote $J_c(\theta) = c(\theta) - \frac{1-F(\theta)}{f(\theta)}c'(\theta)$, which equals $J(\theta)$ if $c(\theta) = \theta$ and is increasing if and only if $J'(\theta)c'(\theta) \geq \frac{1-F(\theta)}{f(\theta)}c''(\theta)$. Analogous to the previous section, for any given cutoff $\hat{\theta} \in [0, \bar{\theta}]$, the revenue maximization problem is equivalent to

$$\max_{s \in \mathrm{MPS}(F) \text{ on } [\hat{\theta},\bar{\theta}]} \int_{\hat{\theta}}^{\bar{\theta}} J_c(\theta)s(\theta)\,\mathrm{d}F(\theta) + \int_{\hat{\theta}}^{\bar{\theta}} \left[(v(\theta) - c(\theta)) - \frac{1-F(\theta)}{f(\theta)}(v'(\theta) - c'(\theta)) - U(\hat{\theta})\right] \mathrm{d}F(\theta) \tag{22}$$

where $s(\theta) = 1 - t(\theta)$ is increasing. Because $U(\theta) \geq 0$ (IR), it is optimal to set $U(\hat{\theta}) = 0$.[15] Assume $J_c(\theta)$ is increasing; then by Proposition 2 in KMS, the solution is $t^*(\theta) = 1 - F(\theta)$ (i.e., $s^*(\theta) = F(\theta)$) for $\theta \geq \hat{\theta}$. Now we solve for the optimal cutoff $\hat{\theta}$.

$$\max_{\hat{\theta} \in [0,\bar{\theta}]} R(\hat{\theta}) = \max_{\hat{\theta} \in [0,\bar{\theta}]} \int_{\hat{\theta}}^{\bar{\theta}} \left[J_c(\theta)F(\theta) + (v(\theta) - c(\theta)) - \frac{1-F(\theta)}{f(\theta)}(v'(\theta) - c'(\theta))\right] \mathrm{d}F(\theta) \tag{23}$$

Denote $J_{vc}(\theta) = (v(\theta) - c(\theta)) - \frac{1-F(\theta)}{f(\theta)}(v'(\theta) - c'(\theta))$. Assume R is concave; the optimal cutoff for revenue maximization, $\hat{\theta}^*$, is given by the first-order condition

$$R'(\hat{\theta}^*) = -(J_c(\hat{\theta}^*)F(\hat{\theta}^*) + J_{vc}(\hat{\theta}^*))f(\hat{\theta}^*) \leq 0, \quad R'(\hat{\theta}^*) \cdot \hat{\theta}^* = 0 \tag{24}$$

Proposition 8. *Assume $c(\theta) - \frac{1-F(\theta)}{f(\theta)}c'(\theta)$ is increasing and R is quasi-concave. The revenue-maximizing mechanism excludes types below $\hat{\theta}^*$ and fully separates types above $\hat{\theta}^*$. In particular, it has no exclusion ($\hat{\theta}^* = 0$) if and only if $v(0) - c(0) - \frac{v'(0)-c'(0)}{f(0)} \geq 0$.*

[15] Because $v(0) - c(0) \geq 0$ and $v'(0) > c'(0)$, $p(\hat{\theta}) = v(\hat{\theta}) - t(\hat{\theta})c(\hat{\theta}) \geq 0$ is satisfied.

4.2 Consumer welfare maximization

The consumer welfare maximization problem is given by

$$\max_{t(\theta),p(\theta),\hat{\theta}} \int_{\hat{\theta}}^{\bar{\theta}} U(\theta)\,\mathrm{d}F(\theta) \tag{25}$$

subject to constraints (17)–(21) on $[\hat{\theta},\bar{\theta}]$.

Analogous to the previous section, for any given cutoff $\hat{\theta}\in[0,\bar{\theta}]$, the welfare maximization problem is equivalent to

$$\max_{s\in\mathrm{MPS}(F)\text{ on }[\hat{\theta},\bar{\theta}]} \int_{\hat{\theta}}^{\bar{\theta}} \left(\frac{1-F(\theta)}{f(\theta)}\right) c'(\theta)s(\theta)\,\mathrm{d}F(\theta) + \int_{\hat{\theta}}^{\bar{\theta}} \left(\frac{1-F(\theta)}{f(\theta)}(v'(\theta)-c'(\theta)) + U(\hat{\theta})\right)\mathrm{d}F(\theta) \tag{26}$$

Assume $\frac{1-F(\theta)}{f(\theta)}c'(\theta)$ is decreasing. Then, the optimal $t^*(\theta)$ pools all types $\theta\in[\hat{\theta},\bar{\theta}]$, that is,

$$t^*(\theta) = 1 - \frac{1}{1-F(\hat{\theta})}\int_{\hat{\theta}}^{\bar{\theta}} F(\theta)\,\mathrm{d}F(\theta) = (1-F(\hat{\theta}))/2. \tag{27}$$

Now we solve for the optimal cutoff $\hat{\theta}$.

$$\max_{\hat{\theta}\in[0,\bar{\theta}]} W(\hat{\theta}) = \max_{\hat{\theta}\in[0,\bar{\theta}]} \int_{\hat{\theta}}^{\bar{\theta}} \frac{1-F(\theta)}{f(\theta)}[v'(\theta) - (1-F(\hat{\theta}))c'(\theta)/2] + U(\hat{\theta})\,\mathrm{d}F(\theta) \tag{28}$$

Note that $W(\hat{\theta})$ is discontinuous at $\hat{\theta}=0$. At $\hat{\theta}=0$, it is optimal to set $p(0)=0$ and thus $U(0)=v(0)-t(0)c(0)=v(0)-c(0)/2\geq 0$, so

$$W(0) = \int_0^{\bar{\theta}} \frac{1-F(\theta)}{f(\theta)}(v'(\theta)-c'(\theta)/2)\,\mathrm{d}F(\theta) + (v(0)-c(0)/2). \tag{29}$$

If $\hat{\theta}>0$, it has to be $U(\hat{\theta})=0$, so the maximal welfare is

$$\max_{\hat{\theta}\in[0,\bar{\theta}]} W(\hat{\theta})|_{U(\hat{\theta})=0} = \max_{\hat{\theta}\in[0,\bar{\theta}]} \int_{\hat{\theta}}^{\bar{\theta}} \frac{1-F(\theta)}{f(\theta)}[v'(\theta)-(1-F(\hat{\theta}))c'(\theta)/2]\,\mathrm{d}F(\theta). \tag{30}$$

We need to compare $W(0)$ and $\max_{\hat{\theta}\in[0,\bar{\theta}]} W(\hat{\theta})|_{U(\hat{\theta})=0}$. If $W(0)\geq\max_{\hat{\theta}\in[0,\bar{\theta}]} W(\hat{\theta})|_{U(\hat{\theta})=0}$, then no exclusion $\hat{\theta}^{**}=0$ is optimal. Otherwise, some exclusion $\hat{\theta}^{**}=\arg\max_{\hat{\theta}\in[0,\bar{\theta}]} W(\hat{\theta})|_{U(\hat{\theta})=0} >$

0 is optimal; the first-order condition is

$$W'(\hat{\theta})|_{U(\hat{\theta})=0} = -(1-F(\hat{\theta}))(v'(\hat{\theta}) - (1-F(\hat{\theta}))c'(\hat{\theta})/2) + f(\hat{\theta})\int_{\hat{\theta}}^{\bar{\theta}} \frac{1-F(\theta)}{2f(\theta)}c'(\theta)\,\mathrm{d}F(\theta) \quad (31)$$

Intuitively, exclusion has a "fixed cost" to consumer welfare as it would involve setting a positive price $p(\hat{\theta}) = v(\hat{\theta}) - t(\hat{\theta})c(\hat{\theta}) \geq 0$. On the other hand, exclusion may also increase consumer welfare by reducing the waiting time for the remaining consumers. Perhaps surprisingly, if $\frac{1-F(\theta)}{f(\theta)}c'(\theta)$ is decreasing (which implies the IFR property), $W'(\hat{\theta}) < 0$ for all $\hat{\theta} \in [0, \bar{\theta}]$, so exclusion always reduces consumer welfare.

Proposition 9. *Assume* $\frac{1-F(\theta)}{f(\theta)}c'(\theta)$ *is decreasing (which implies the IFR property). The consumer welfare-maximizing mechanism has no exclusion and offers one priority level to all types.*

5 Discussions

With a feasibility condition in Theorem 1, the mechanism design approach can be applied to allocation problems in which allocating the object to one agent has externalities on other agents, such as waiting time (Gershkov and Winter, 2023), status (Moldovanu et al., 2007), and conspicuous goods (Rayo, 2013).

By applying this approach to GW's model, I study the effect of PS on the monopoly provider's revenue and consumer welfare and provide the necessary and sufficient conditions for some of their propositions while allowing for stochastic priority levels. In particular, I show that the trade-off between the provider's revenue and consumer welfare is less stark. Although an increasing (decreasing) failure rate implies that adding more priority levels increases (may decrease) the revenue but decreases (increases) consumer welfare at the same time, it is not a necessary condition. Thus, if the cost distribution has a decreasing failure rate but satisfies Myerson's regularity condition (e.g., Pareto distribution), the two objectives are aligned—adding more priority levels will increase *both* the revenue and consumer welfare. An all-pay auction among customers can implement full separation. Moreover, the provider can guarantee at least half the maximal revenue by offering two priority levels. The approximation can be arbitrarily close if the cost distribution is sufficiently concave.

References

Border, Kim C. (1991), "Implementation of Reduced Form Auctions: A Geometric Approach." *Econometrica*, 59, 1175–1187. 10.2307/2938181. [4]

Börgers, Tilman (2015), *An Introduction to the Theory of Mechanism Design*, New York, NY: Oxford University Press. [7]

Gershkov, Alex and Paul Schweinzer (2010), "When Queueing Is Better than Push and Shove." *International Journal of Game Theory*, 39, 409–430. 10.1007/s00182-009-0198-x. [8]

Gershkov, Alex and Eyal Winter (2023), "Gainers and Losers in Priority Services." *Journal of Political Economy*, 131, 3103–3155. 10.1086/724850. [1, 14]

Kleiner, Andreas, Benny Moldovanu, and Philipp Strack (2021), "Extreme Points and Majorization: Economic Applications." *Econometrica*, 89, 1557–1593. 10.3982/ECTA18312. [2, 4]

Maskin, Eric and John Riley (1984), "Optimal Auctions with Risk Averse Buyers." *Econometrica*, 52, 1473–1518. 10.2307/1913516. [4]

Matthews, Steven A. (1984), "On the Implementability of Reduced Form Auctions." *Econometrica*, 52, 1519–1522. 10.2307/1913517. [4]

Moldovanu, Benny, Aner Sela, and Xianwen Shi (2007), "Contests for Status." *Journal of Political Economy*, 115, 338–363. 10.1086/518010. [1, 2, 14]

Mussa, Michael and Sherwin Rosen (1978), "Monopoly and Product Quality." *Journal of Economic Theory*, 18, 301–317. 10.1016/0022-0531(78)90085-6. [10]

Onuchic, Paula and Debraj Ray (2023), "Conveying Value via Categories." *Theoretical Economics*, 18, 1407–1439. 10.3982/TE5026. [3]

Rayo, Luis (2013), "Monopolistic Signal Provision." *The B.E. Journal of Theoretical Economics*, 13, 27–58. 10.1515/bejte-2012-0003. [3, 14]

Xiao, Peiran (2024), "Incentivizing Agents through Ratings." July, arXiv:2407.10525. [3]