Fair and Efficient Allocation with Quotas

Siddhartha Banerjee SBANERJEE@CORNELL.EDU
Matthew Eichhorn MAE226@CORNELL.EDU
Cornell University

David Kempe DAVID.M.KEMPE@GMAIL.COM
University of Southern California

Abstract

In many settings, such as the rationing of medical care and supplies, university admissions, and the assignment of public housing, the decision of who receives an allocation can be justified by various normative criteria (ethical, financial, legal, etc.). These criteria can influence priorities and restrict the number of units available to particular demographics. We consider a setting wherein a set of identical items must be distributed among unit-demand applicants. The items are divided across a set of categories (each with an allocation quota), and each category has a priority ordering over its eligible applicants. Building on previous work, we define a set of natural desiderata for allocations. We use techniques from linear and integer programming to give a polynomial-time procedure for finding *Pareto efficient* allocations satisfying these desiderata. The clean formulation of our algorithm allows us to more easily derive known results in this space and provides the flexibility to consider alternate objectives and enforce various notions of equity between categories.

1. Introduction

The way in which a society distributes its resources can have profound impacts on the population. While not new, this fact has been forcefully reintroduced into public consciousness by the COVID-19 pandemic. Scarcities brought about by the pandemic have caused researchers in many fields — including doctors, ethicists, psychologists, and economists — to question how we allocate care (White and Lo, 2020; Andrews et al., 2021; Emanuel et al., 2020; Binkley and Kemp, 2020; Pathak et al., 2021). Finding a good allocation is far from straightforward, as legal, financial, and ethical considerations can place nuanced requirements on the set of allowable allocations. As examples of such restrictions, consider the following settings:

Academic Fellowships: Alumni benefactors of an institution may establish scholarship funds to encourage the enrollment of students with certain demographics, backgrounds, skills, etc. To make use of these funds, the university must admit students who meet the qualifications of these awards. We use this as our motivating example throughout.

Medical Care: To ensure an equitable distribution of vaccines, the COVAX program has set standards for the distribution of vaccines to developing countries, as well as for the prioritization of vaccinating target groups including healthcare workers, the elderly, and individuals with comorbidities (COVAX). This application is discussed in detail by Pathak et al. (2021).

Primary School Enrollment: In Boston, half of each school's seats are reserved for students in its neighborhood, and a school must give priority to students with siblings also attending that school (Abdulkadiroğlu et al., 2005). In Chicago, a school integration policy requires that magnet schools allocate roughly one fourth of their seats to students in each of four socio-economic tiers (Benabbou et al., 2019). Chile's School Inclusion Law defines which factors can and cannot be used to prioritize students, and ensures equitable admission to students with economic hardships (Correa et al., 2021).

Public Housing: In Singapore, a 1989 Ethnic Integration Policy places a quota on the number of units in each public housing development that may be allocated to each of three major ethnic groups: Chinese, Malay, and Indian/Other (Benabbou et al., 2019, 2018).

We consider a model of reserve systems introduced by Pathak et al. (2021), which captures a set of common features shared by all of the settings above to varying extents. The formal model is given in Section 2, but we briefly summarize these features below:

Rationing and Quotas: The available resource is less than the demand, and so must be rationed. To implement the rationing, the total resource budget is split into a quota q_c for each of multiple allocation categories $c \in \mathcal{C}$. For example, in the case of academic fellowships, quotas could be reserved for local students, under-represented minorities, international students, family members of alumni, and a general pool.

Eligibility and Priority Rules: Each category has criteria for determining the eligibility and priority of each agent. In our model, this corresponds to each category having a preorder over a subset of students who meet the eligibility criterion for that category. Agents may be eligible for multiple categories, so categories may need to coordinate allocations. Importantly, there may be no natural way to compare agents with the same rank in different categories. Note also that the eligibility may be defined independently of the quota, and thus may lead to unavoidable wastage in any allocation.

Agent Indifference: Each agent wants a single unit of the resource, but they are indifferent as to which category awards them the resource[1]. More generally, agents can get allocations in $[0, 1]$ from multiple categories, such that the total allocation is bounded by 1; agents' utilities are assumed to be non-decreasing in their total allocation. For example, in the case of academic fellowships, a student can receive partial support from different funding sources, with a cap on the total award.

Pathak et al. (2021) argue that this setting captures many critical features of quota-based rationing. It emphasizes the role of categories and quotas as enforcing incomparable norms, rather than maximizing some explicit utility. On the other hand, agent indifference is tautological for resources like vaccines/medical supplies/scholarships; it also holds to an extent for (limited) resources like school seats and housing, where the difference in utility from receiving vs. not receiving an allocation vastly outweighs inter-category utility differences.

The core question now is what features are desired for an allocation in such a setting. To this end, Pathak et al. (2021) posit a set of three 'axiomatic' requirements. The first two make natural

1. We note that this feature may not be true in all of the examples. For example, families are likely to have preferences about which schools their students attend and the location of their housing.

impositions that each category allocates as much as possible while respecting its quota, and also allocates only to eligible agents; these are standard and easily implemented. The third axiom ensures that priorities are respected, by requiring that a category never allocate to an agent if a higher-priority agent does not have full allocation; this is a trickier requirement, in particular, when combined with partial eligibility lists. Indeed, past work on this problem (Pathak et al., 2021; Delacrétaz, 2021) eschews finding Pareto efficient allocations, and while Aziz and Brandl (2021) provide a scheme to find a maximal (and hence Pareto efficient) allocation, they do not give any insight into the structure of satisfactory allocations, or how one can select from them to improve any secondary performance measure such as equity of allocations.

1.1. Our Contributions

Our work considers the setting described above (and formalized in Section 2), and attempts to provide a simple characterization and computational procedure for finding *satisfactory allocations*: those obeying the main axioms of previous work (respect for eligibility, quotas, and priority), but in addition, guaranteeing *Pareto efficiency* from the viewpoint of agents, and *stability* from the viewpoint of categories. To this end, our main result (in Section 3) shows that *satisfactory allocations can be realized as the solutions to a simple weighted matching LP*. In particular, our characterization uses a technique from integer programming whereby we perturb the unweighted matching LP while ensuring that its solutions are satisfactory allocations. Moreover, we show that in fact *all satisfactory integer allocations can be realized as optimal matchings under some appropriate perturbation* of edge weights.

Intuitively, our use of perturbations allow us to transform the problem with (partial) cardinal preferences to an ordinal welfare maximization problem. This reformulation allows us not only to more easily derive known results, but enables several new insights and extensions. In Section 4, we show that our perturbed LP approach leads to allocations that satisfy additional desiderata, which are natural for this setting, but are violated by existing algorithms. Next, our approach admits a simple extension in settings where quotas are not binding, and some units can be de-reserved by categories to increase overall welfare. In terms of insights, we show that *every* Pareto optimal solution turns out to allocate the same number of units, which moreover equals the optimal matching *without* priority requirements. We also give an efficient procedure to identify agents who are present in all satisfactory allocations. Finally (and most importantly), our approach gives new tools for selecting satisfactory allocations to achieve some secondary objective — in Section 5, we describe how we can use perturbations to optimize two different notions of equity in our setting.

1.2. Related Work

As mentioned, we build on the recent work of Pathak et al. (2021), which has also inspired several other follow-up papers, with two of particular note. On the question of allocation selection, Delacrétaz (2021) notes that the policy of Pathak et al. (2021) is not uniquely specified, and different choices can induce biases in the allocation. He then attempts to allay this concern by introducing a waterfilling-style *simultaneous allocation* procedure that leads to a unique (fractional) outcome. On the other hand, Aziz and Brandl (2021) introduce a procedure that ensures a maximum-size allocation, even with partial eligibility. We discuss these policies, and highlight some of their shortcomings in Appendix A. In the interest of space, we refer the reader to the thorough discussion of related models and practical applications of this setting presented in these papers.

A closely related problem to reserve allocation is fair division, where agents have preferences over (non-identical) items, and we seek a Pareto efficient division. There are key distinctions between these models with regard to notions of *stability* and *utility*: in fair division, both the preferences that determine the stability and efficiency of a solution belong to agents; however, in reserve allocation, stability is dictated by category preferences while utility is dictated by agent allocations. Despite these semantic differences, the structure of desired allocations in both turn out to be quite similar[2]. In particular, Saban and Sethuraman (2015) consider the problem of determining the probability of a match under random serial dictatorship, and describe a polynomial-time procedure for locating "necessary" (agent, object) pairs which is similar to our discussion of unanimous agents (Section 4.3). Moreover, Biró and Gudmundsson (2021) propose using welfare maximization to compute Pareto efficient fair division solutions. The perturbation of the b-matching polytope that we discuss in Section 3 serves as a unifying viewpoint for these problems.

Finally, settings with two-sided preferences have a long history, stemming from Gale and Shapley's seminal work on the deferred acceptance (DA) algorithm (Gale and Shapley, 1962). While a fairly robust algorithm, DA can fail to compute a Pareto efficient allocation in the case of indifferences, as pointed out by Erdil and Ergin (2017). They describe an iterative procedure to Pareto improve an allocation while preserving its stability, illustrating that notions of stability and efficiency can be simultaneously realized. The flow-augmentation ideas in their improvement procedure share commonalities with our arguments in Section 3.

2. Model

2.1. Basic Setting

Resources, Categories, and Agents: A set $\mathcal{A}$ of n *agents* compete for q indistinguishable, indivisible units of a resource (admission, residence, vaccine, etc.). The units are distributed to a set $\mathcal{C}$ of m *categories*, through which they are allocated. Each category $c \in \mathcal{C}$ is given a *quota* of q_c units to allocate, such that $q = \sum_c q_c$. Each agent wants one unit of the resource but is indifferent as to which category provides their allocation.

Eligibility and Priorities: Each category partitions $\mathcal{A}$ into a set of *eligible* and *ineligible* agents. The eligible agents are further partitioned into *priority* tiers.

Formally, each category $c \in \mathcal{C}$ has a total preorder $\succeq_c$ over $\mathcal{A} \cup \{\theta_c\}$, where θ_c is an additional symbol used to represent the eligibility threshold in category c. Given any two agents $a, a' \in \mathcal{A}$, $a \succeq_c a'$ denotes that a has weakly higher priority than a' in c. We write $a \sim_c a'$ when a and a' have the same priority in c, i.e., when $a \succeq_c a'$ and $a' \succeq_c a$; we write $a \succ_c a'$ when $a \succeq_c a'$ and $a' \not\succeq_c a$, so a has (strictly) higher priority in c. We assume that $\theta_c \not\sim_c a$ for any $a \in \mathcal{A}$, and interpret the eligible agents in c as those a with $a \succ_c \theta_c$. Given any agent a and any category c, we define the *ranking* of a in c, denoted by $r_c(a)$, to be the length ℓ of the longest chain $a_1 \succ_c a_2 \succ_c \cdots \succ_c a_\ell = a$ with each $a_i \in \mathcal{A}$. Note that $1 \leq r_c(a) \leq n$.

We visualize category quotas/priorities/eligibility using charts in the style of Fig. 1.

Allocations: Our goal is to find an *allocation* $\mathbf{x} : \mathcal{A} \times \mathcal{C} \to [0, 1]$. We interpret $x_{a,c}$ as the probability that agent a receives an allocation from category c; the desired marginal probabilities can be realized via a standard Birkhoff-von Neumann decomposition of fractional allocations as

2. Indeed, our results provide some intuition as to why this is the case, as when viewed as an ordinal welfare maximization problem, it is clear that the two sides of the market are symmetric.

α (1)	β (1)	γ (1)
c	a	b , c
	b	a

Figure 1: An instance with $\mathcal{C} = \{\alpha, \beta, \gamma\}$ with quotas $(1, 1, 1)$, and $\mathcal{A} = \{a, b, c\}$. In each category c, the agents listed in the i'th row have ranking $r_c(\cdot) = i$, and the threshold θ_c is just below the last listed agent (that is, only eligible agents are listed).

a convex combination of integral matchings. Note that our unit-demand assumption allows us to restrict attention to the case where $\sum_c x_{a,c} \leq 1$ for each $a \in \mathcal{A}$. Note also that when each $x_{a,c} \in \{0, 1\}$, $\mathbf{x}$ coincides with an allocation map $\varphi : \mathcal{A} \to \mathcal{C} \cup \{\varnothing\}$.

2.2. Primary Desiderata for Satisfactory Allocations

A natural question now, given the above setting, is which properties make an allocation "satisfactory." The following two desiderata were proposed as axioms by Pathak et al. (2021) (and adopted by Delacrétaz (2021) and Aziz and Brandl (2021)) as a natural formalism implied by the agent eligibility and priorities. We inherit the statements for fractional matchings from Delacrétaz, as these naturally generalize their integral counterparts.

[ER] Eligibility Respecting: No agent receives any allocation through a category for which they are ineligible. Formally, we have

$$x_{a,c} > 0 \implies a \succ_c \theta_c \quad \text{for all } a \in \mathcal{A}, c \in \mathcal{C}.$$

[PR] Priority Respecting: If an agent a receives any allocation through a category c, then each agent a' with higher priority in c is fully allocated. Formally

$$x_{a,c} > 0 \ \wedge \ a' \succ_c a \implies \sum_{c' \in \mathcal{C}} x_{a',c'} = 1 \quad \text{for all } a \in \mathcal{A}, c \in \mathcal{C}.$$

Another axiom states that no category's allocation exceeds its quota. (This is either explicitly stated or implied in (Pathak et al., 2021; Delacrétaz, 2021; Aziz and Brandl, 2021).)

[QR] Quota Respecting: The total allocation from category c is at most q_c. Formally

$$\sum_{a \in \mathcal{A}} x_{a,c} \leq q_c \quad \text{for all } c \in \mathcal{C}.$$

We note that this is a somewhat restrictive assumption, and it may lead to inefficiencies due to incomplete allocations. Indeed, the idea of *de-reserving* unfilled quotas has been considered with regard to affirmative action policies in India (Aygun and Turhan, 2021) (for a thorough introduction on this subject, see (Sönmez et al., 2019)). In Section 4.2, we introduce an alternate setting that allows for some de-reservation of quotas if it results in more allocation.

Henceforth, unless stated otherwise, any allocation we consider is assumed to satisfy **[ER]**, **[PR]** and **[QR]**. Next, from the agents' standpoint, a natural desideratum for an allocation is that it be Pareto efficient.

[PE] Pareto Efficient: There is no alternate allocation $\mathbf{y}$ satisfying **[ER]**, **[QR]**, **[PR]** in which one agent gets a strictly higher allocation and no agent receives a lower allocation.

$$\text{there is an } a \in \mathcal{A} \ : \sum_{c \in \mathcal{C}} y_{a,c} > \sum_{c \in \mathcal{C}} x_{a,c} \implies \text{there is an } \ a' \in \mathcal{A} \ : \sum_{c \in \mathcal{C}} y_{a',c} < \sum_{c \in \mathcal{C}} x_{a',c}.$$

While **[PE]** is clearly a desirable property, it is a stronger efficiency requirement than those considered in prior work; in particular, it implies the "*non-wastefulness*" axiom used in (Pathak et al., 2021; Delacrétaz, 2021; Aziz and Brandl, 2021). Although the "*maximum size matching*" property of Aziz and Brandl (2021) appears to be stronger than Pareto efficiency, we argue that they are in fact equivalent in Section 3.1. We find Pareto efficiency to be a more natural desideratum.

In contrast to two-sided matching settings, we only consider **[PE]** from the point of the agents — this builds on the idea that category priorities and eligibility are used in such settings to implement normative criteria, rather than having associated utilitarian implications. Nevertheless, from the point of view of implementation/interpretation, it may still be useful to consider when an allocation can be considered satisfactory from the viewpoint of the categories. Our final primary desideratum addresses this point.

[S] Stable: There is no way for categories to transfer allocation to agents of higher priority. More formally, there do not exist agents $a_0, a_1, \ldots, a_j = a_0$ and categories $c_0, c_1, \ldots, c_j = c_0$ such that $x_{a_i,c_i} > 0$ and $a_{i+1} \succ_{c_i} a_i$ for each $0 \leq i < j$.

This property had not been considered in the earlier literature; we argue that it is natural in settings where the categories can derive some secondary utility from their allocated agents. As a concrete example, consider a university or funding agency awarding fellowships to students, where each category corresponds to a donor who has established specific criteria for who should be awarded from their donated funds. Such donors may want to advertise their awardees, and thus want them to reflect their criteria as much as possible.

In this work, we consider the above to constitute the *primary desiderata* for any allocation

Definition 1 (Satisfactory Allocation) *An allocation is* satisfactory *if it satisfies all of the primary desiderata:* **[ER]**, **[PR]**, **[QR]**, **[PE]**, *and* **[S]**.

Fig. 2 depicts some allocations of the instance from Fig. 1, and discusses which desiderata they satisfy/violate. We note again that while the first three have all been considered in (Pathak et al., 2021; Delacrétaz, 2021; Aziz and Brandl, 2021), **[PE]** is only proposed as being desirable but not implemented in (Pathak et al., 2021; Delacrétaz, 2021) (and is indirectly considered in (Aziz and Brandl, 2021)), while **[S]** is not considered in any of these works.

2.3. Additional Properties of Allocation Rules

In light of the above, the foremost property of any allocation rule is that in any given instance, it returns a satisfactory allocation (i.e., one which obeys the above primary desiderata). There are several additional features that one could desire from allocation rules in our setting. While we do not *a priori* impose that these be satisfied by a rule under consideration, they turn out to be immediate consequences of our approach.

The first property we consider reflects a natural desire that under a given rule, agents need not be un-allocated when any category increases its quota (holding all else the same). We consider

Allocation 1

α (1)	β (1)	γ (1)
$\boxed{c}$	$\boxed{a}$	$\boxed{b}$, c
	b	a

Allocation 2

α (1)	β (1)	γ (1)
c	a	b, $\boxed{c}$
	$\boxed{b}$	a

Allocation 3

α (1)	β (1)	γ (1)
c	$\boxed{a}$	b, $\boxed{c}$
	b	a

Allocation 4

α (1)	β (1)	γ (1)
c	a	b, c
	$\boxed{b}$	$\boxed{a}$

Figure 2: Four (integer) allocations of the instance from Fig. 1. Allocation 1 is the (unique) satisfactory allocation in this instance. Allocation 2 violates **[PR]**: b is allocated in category β, but a, who has higher priority, remains unallocated. Allocation 3 violates **[PE]**, as it is Pareto dominated by Allocation 1. Allocation 4 violates **[S]**: categories β and γ can switch and each allocate to a higher-priority agent.

possibly non-deterministic rules ψ mapping instances to non-empty sets of allocations, and define $\psi(I)$ to be the set of allocations that could be produced on input I.

Monotonicity [M]: The allocation rule ψ is *monotone* if (and only if) the following holds for all pairs of matching instances I, I' with the same $\mathcal{A}, \mathcal{C}, \{\succeq_c\}$, and with $q'_c \geq q_c$ for each c: For every $\mathbf{x} \in \psi(I)$, either $\mathbf{x} \in \psi(I')$, or there exists an allocation $\mathbf{y} \in \psi(I')$ which Pareto dominates $\mathbf{x}$.

Next, we turn to strategic considerations. Ideally we would like our allocation rule to be resilient to misreporting of priorities/eligibility. However, Example 1 shows that categories can improve their outcome through strategic manipulation.

Example 1 *Consider the following allocation instance:*

α (1)	β (1)
a	*a*
b	*c*

While any satisfactory allocation must fully allocate to a, the remaining unit can be arbitrarily divided among b and c. Knowing β's priority list, α can choose to declare a ineligible, which leads to both of α's eligible agents receiving a full allocation: a through β and b through α.

A weaker form of strategic behavior by categories (that turns out is possible to disincentivize) is one where a category subdivides and reapportions its quotas. The Sybil-proofness property requires that such manipulation does not give categories a strategic advantage.

Sybil-proofness [SP]: No category c can split into multiple categories $c_1, \ldots, c_k$, each with identical priority lists to c and with $\sum_{i \in [k]} q_{c_i} = q_c$, in a way that increases the total allocation of their eligible agents. Similarly, categories with identical preference lists cannot merge to increase the allocation of their eligible agents.

From the vantage point of the agents, a natural notion of strategyproofness proposed by Aziz and Brandl (2021) is that no agent benefits by intentionally worsening their ranking (for example, by scoring poorly on a placement exam).

Strategyproofness [ST]: No agent can receive a greater allocation by decreasing their priority in a category.

Finally, an important desideratum, in particular for large instances, is that the allocation rule be efficiently computable.

Computational Efficiency [CE]: The allocation rule produces a satisfactory allocation in time polynomial in m, n, and q.

In Appendix A, we outline the existing algorithms for this problem (Pathak et al., 2021; Delacrétaz, 2021; Aziz and Brandl, 2021), and discuss ways in which they can violate our desiderata. On the other hand, in Section 4.1, we discuss how the class of policies we propose next naturally satisfies all of them.

3. Satisfactory Allocations via Linear Programming

In this section, we give our main result: we show that for any given instance, a satisfactory allocation can be found using a simple linear program. Thereto, note first that the unit demand of each agent, as well as the **[ER]** and **[QR]** constraints, together can be encoded as a b-matching polytope. Any maximizer of the total allocation $V(\mathbf{x}) := \sum_{a\in\mathcal{A}}\sum_{c\in\mathcal{C}} x_{a,c}$ (satisfying the other desiderata) is Pareto efficient. Therefore, we can enforce Pareto efficiency by taking $V(\mathbf{x})$ as our objective. Later, we argue that our consideration of maximal allocations is without restriction. Overall, this gives the following b-matching LP:

$$
\begin{aligned}
(P) \qquad \max \quad & V(\mathbf{x}) && \\
\text{subject to} \quad & \sum_{a\in\mathcal{A}} x_{a,c} \le q_c && \text{for all } \ c\in\mathcal{C} \\
& \sum_{c\in\mathcal{C}} x_{a,c} \le 1 && \text{for all } \ a\in\mathcal{A} \\
& x_{a,c} = 0 && \text{for all } \ a\in\mathcal{A}, c\in\mathcal{C} \text{ with } \theta_c \succ_c a \\
& x_{a,c} \ge 0 && \text{for all } \ a\in\mathcal{A}, c\in\mathcal{C}.
\end{aligned}
$$

As written, (P) does not express any notion of respect for priorities or stability. However, while encoding the feasible set under these constraints appears non-trivial, the critical observation is that we can perturb the coefficient of each $x_{a,c}$ in the objective to $1-\delta_{a,c}$ in such a way as to ensure that *any optimal solution* to the perturbed LP satisfies both of these desiderata. To do this, we introduce the notion of a *valid perturbation*.

Definition 2 (Valid Perturbation Profile) *A perturbation profile $(\delta_{a,c})_{a\in\mathcal{A},c\in\mathcal{C}}$ is* valid *if it has the following three properties:*

Positivity: *Each* $\delta_{a,c} > 0$.

Small Effect: $\sum_{a\in\mathcal{A}}\sum_{c\in\mathcal{C}} \delta_{a,c} < \frac{1}{3}$.

Consistency: *$a \succeq_c a'$ if and only if $\delta_{a',c} \geq \delta_{a,c}$*

Now, we consider the modified objective

$$V_\delta(\mathbf{x}) := \sum_{a \in \mathcal{A}} \sum_{c \in \mathcal{C}} x_{a,c}\big(1 - \delta_{a,c}\big) = V(\mathbf{x}) - \sum_{a \in \mathcal{A}} \sum_{c \in \mathcal{C}} \delta_{a,c}\, x_{a,c}.$$

We let (P_δ) be the linear program with the same constraint polytope as (P), but with objective $V_\delta(\mathbf{x})$. The following theorem shows that the solutions to these perturbed linear programs give allocations satisfying all of our desiderata.

Theorem 3 *Let δ be any valid perturbation profile, and let $\mathbf{x}^*$ be a solution to (P_δ). Then, $\mathbf{x}^*$ is a satisfactory allocation.*

Proof The constraints immediately ensure that any feasible solution of (P_δ) satisfies **[ER]** and **[QR]**. To establish **[PR]**, let $\mathbf{x}$ be a feasible solution, a, a' be agents and c a category such that $a' \succ_c a$, $x_{a,c} = \varepsilon_1 > 0$ and $\sum_{c'} x_{a',c'} = 1 - \varepsilon_2 < 1$. Then, we can decrease $x_{a,c}$ and increase $x_{a',c}$ by $\min(\varepsilon_1, \varepsilon_2)$ without violating any constraints. Since δ is consistent, we have $\delta_{a,c} < \delta_{a',c}$, so the reassignment strictly increases the objective value. Thus, such an $\mathbf{x}$ is not optimal, and $\mathbf{x}^*$, being optimal, satisfies **[PR]**.

To establish **[S]**, for a feasible LP solution $\mathbf{x}$, let $a_1, \ldots, a_j = a_0$ be agents and $c_1, \ldots c_j = c_0$ be categories with $x_{a_i,c_i} > 0$ and $a_{i+1} \succ_{c_i} a_i$ for each $0 \leq i < j$. By the unit demand constraints, note that each $x_{c_i,a_{i+1}} < 1$, so we may define

$$\varepsilon := \min\Big\{ \min_{i=0,\ldots,j-1} \big\{x_{a_i,c_i}\big\}, \min_{i=0,\ldots,j-1} \big\{1 - x_{a_{i+1},c_i}\big\} \Big\} > 0.$$

Let $\mathbf{y}$ be the allocation obtained by decreasing each x_{a_i,c_i} by ε and increasing each x_{a_{i+1},c_i} by ε. Note that $\mathbf{y}$ remains feasible in (P_δ). Since δ is consistent, we have $\delta_{a_{i+1},c_i} < \delta_{a_i,c_i}$, for each $0 \leq i < j$. Therefore,

$$V_\delta(\mathbf{y}) - V_\delta(\mathbf{x}) = \varepsilon \cdot \sum_{i=0}^{j-1} (\delta_{a_i,c_i} - \delta_{a_{i+1},c_i}) > 0.$$

Again, such an $\mathbf{x}$ cannot be optimal; hence, $\mathbf{x}^*$, being optimal, satisfies **[S]**.

It remains to establish **[PE]**. Note that for any optimal solution $\hat{\mathbf{x}}$ to (P), we have

$$V(\mathbf{x}^*) \geq V_\delta(\mathbf{x}^*) \geq V_\delta(\hat{\mathbf{x}}) = V(\hat{\mathbf{x}}) - \sum_{a \in \mathcal{A}} \sum_{c \in \mathcal{C}} \hat{x}_{a,c}\delta_{a,c} \geq V(\hat{\mathbf{x}}) - \sum_{a \in \mathcal{A}} \sum_{c \in \mathcal{C}} \delta_{a,c} > V(\hat{\mathbf{x}}) - \frac{1}{3}.$$

Here, the first inequality follows since each $x^*_{a,c}, \delta_{a,c} \geq 0$. The second inequality follows since $\mathbf{x}^*$ is an optimal solution to (P_δ). The third follows because the unit demand constraints ensure that each $\hat{x}_{a,c} \leq 1$. Finally, the fourth inequality follows since δ has small effect.

Additionally, $\hat{\mathbf{x}}$ maximizes V among all feasible solutions to (P), which include $\mathbf{x}^*$. Therefore, $V(\hat{\mathbf{x}}) \geq V(\mathbf{x}^*)$. Combining both inequalities, we find that

$$V(\hat{\mathbf{x}}) \geq V(\mathbf{x}^*) > V(\hat{\mathbf{x}}) - \tfrac{1}{3}. \tag{1}$$

Observe that the constraint matrix of (P) is *totally unimodular*, as it encodes a b-matching polytope. Consequently, as long as all the quotas q_c are integral, every corner point of the constraint polytope is integral. Because $V(\mathbf{x})$ is simply a sum of entries $x_{a,c}$, it must be integral at corner points, and therefore at all maximizers $\mathbf{x}$ of V. In particular, because $\hat{\mathbf{x}}$ maximizes V, $V(\hat{\mathbf{x}})$ is integral. If $\mathbf{x}^*$ is a corner point, then $V(x^*)$ is also integral. However, integral solutions satisfying the bounds in Eq. (1) require $V(\hat{\mathbf{x}}) = V(\mathbf{x}^*)$.

If $\mathbf{x}^*$ is not a corner point, then we write $\mathbf{x}^* = \sum_i \lambda_i \mathbf{x}^{(i)}$ as a convex combination of corner points $\mathbf{x}^{(i)}$. Because $\mathbf{x}^*$ maximizes V_δ, each of the $\mathbf{x}^{(i)}$ must also maximize V_δ. By the argument from the previous paragraph, $V(\hat{\mathbf{x}}) = V(\mathbf{x}^{(i)})$ for all i. But then, the convex combination $\mathbf{x}^*$ must also have $V(\mathbf{x}^*) = V(\hat{\mathbf{x}})$. Thus, each maximizer $\mathbf{x}^*$ of V_δ (whether or not it is a corner point) is also a maximizer of V, and hence satisfies **[PE]**. ■

A surprising immediate consequence of this result is that the **[PE]** and **[S]** desiderata are essentially enforceable "for free" (i.e., without loss to the total allocation size).

Corollary 4 *Given quotas $(q_c)_{c\in C}$, let V^* denote the size of the maximum allocation returned by (P) (i.e., satisfying* **[ER]** *and* **[QR]***). Then, for* any *priority orders $(\succeq_c)_{c\in\mathcal{C}}$, there is an integral allocation with total allocation V^* that additionally satisfies* **[PR]** *and* **[S]**.

In other words, one need not compromise on the efficiency of the solution in order to ensure its stability and accommodation of priorities.

We note that the above property is implied by Aziz and Brandl (2021) based on the properties of Algorithm 4. However, Theorem 3 gives a much simpler way to see why this holds. Moreover, it provides a much simpler computational tool for selecting a satisfying allocation: compared to Algorithm 4, which requires one to solve n separate b-matching problems, our approach requires solving a single weighted b-matching problem, which can be efficiently solved, for instance using the Hungarian algorithm (Ramshaw and Tarjan, 2012).

Corollary 5 *A satisfactory allocation can be computed in $O\big(mnV^* + (V^*)^2 \log(\min(n,q))\big) = O(mnq + q^2 \log q)$ time.*

3.1. Attaining all Satisfactory Allocations

By Theorem 3, we know that solving (P_δ) with any valid δ will produce an allocation satisfying all of our desiderata (i.e., a *satisfactory* allocation). A follow-up question is whether *all* (integer) satisfactory allocations are solutions of (P_δ) for some choice of δ. Here, we answer this question in the affirmative. To do so, we give an alternate characterization of the integral satisfactory allocations. Then, for each allocation $\mathbf{x}$ in this alternate characterization, we produce an assignment of δ such that $\mathbf{x}$ is a solution to (P_δ).

One immediate point of concern is that since any solution to (P_δ) maximizes $V(\mathbf{x})$, our LP formulation could miss out on some satisfactory solutions which are non-maximal. Fortunately, this turns out not to be the case.

Lemma 6 *Any allocation $\mathbf{x}$ satisfying* **[PE]** *maximizes $V(\mathbf{x})$.*

Proof We will argue the contrapositive — i.e., any x that does not maximize $V(x)$ is not **[PE]**. Consider the flow network representation of the allocation problem shown in Fig. 3. We color some agent nodes red as follows:

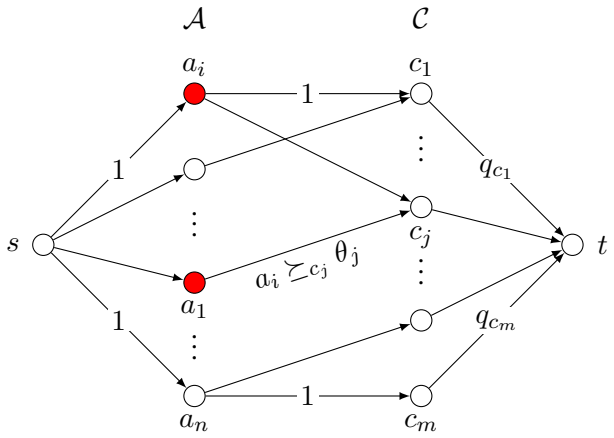

Figure 3: A flow network representation of an allocation instance. The source node s has a unit-capacity edge to each agent node. Each category node has an edge to the sink node t with capacity equal to that category's quota. There are unit-capacity edges from each agent node to the nodes of categories in which it is eligible.

Under the given allocation, for any category c that has an eligible agent who is not fully allocated in $\mathbf{x}$, color all agents eligible in c red (even if they are fully allocated).

If $\mathbf{x}$ is not a maximal allocation, then there is an augmenting path $P = (s, a_1, c_1, \ldots, a_k, c_k, t)$ in this flow network. We record the following observations.

1. a_1 is red: The in-weight of each agent node is its allocation. Augmenting along P will increase the in-weight of its first agent node, so this agent node must not have been fully allocated.

2. c_k has not exhausted its quota: The out-weight of each category node is its allocated quota. Augmenting along P will increase the out-weight of its last category node, so this category must not have exhausted its quota.

3. Given any red agent a, there is a path of the form $s \to a_0 \to c_0 \to a$ in the residual graph for $\mathbf{x}$, where a_0 is a highest-priority agent in c_0 that is not fully allocated: this follows from the definition of red agent nodes.

Let a_i be the *last* red node in P (there must be such a node by Observation 1), and consider the alternate augmenting path $P' = (s, a_0, c_0, a_i, c_i, \ldots, a_k, c_k, t)$ using the "shortcut" from Observation 3. Augmenting along P' will strictly increase the allocation to a_0 and conserves the allocations of $a_i, \ldots, a_k$. Let $\mathbf{y}$ be the allocation after this augmentation. By the construction of the flow network, $\mathbf{y}$ still satisfies **[ER]** and **[QR]**. Moreover, every agent $a \succeq_{c_0} a_0$ is fully allocated, and every agent $a_i, \ldots, a_k$ maintains its allocation in $\mathbf{y}$, so $\mathbf{y}$ also satisfies **[PR]**. Thus, $\mathbf{y}$ is a Pareto improvement to $\mathbf{x}$, meaning $\mathbf{x}$ did not satisfy **[PE]**. ■

Let Σ be the collection of all multiset orderings of $\left\{c^{q_c}\right\}_{c \in \mathcal{C}}$ (i.e., the set of all sequences of length q wherein each category $c \in \mathcal{C}$ appears q_c times). We refer to Σ as the set of *choice orders* for

our reserve system. For a given choice order $\sigma \in \Sigma$, we define the *serial dictatorship allocation* $\mathbf{x}_\sigma$ to be the (integer) allocation[3] obtained by the procedure in Algorithm 1. Note that serial dictatorship generalizes the sequential reserve matchings of Pathak et al. (2021).

Algorithm 1 Serial Dictatorship Allocation

Input: Choice order $\sigma \in \Sigma$

for each $\sigma_i = c$ in σ in order **do**
  if c has remaining quota and an eligible unallocated agent **then**
    c allocates to its highest-priority unallocated agent

Lemma 7 *For all $\sigma \in \Sigma$, the serial dictatorship allocation $\mathbf{x}_\sigma$ satisfies* **[ER]**, **[QR]**, **[PR]**, *and* **[S]**.

Proof It is immediate from the definition of serial dictatorship that $\mathbf{x}_\sigma$ satisfies **[ER]**, **[QR]**, and **[PR]**: each category c only allocates to eligible agents, can allocate to at most q_c agents (σ contains q_c copies of c), and always allocates to a highest-priority unallocated agent. It remains to argue that $\mathbf{x}$ is stable. In any subset S of allocated agents, consider the first time that one agent a was allocated by a category c. By definition, c selected a highest-priority unallocated agent, so $a \succeq_c s$ for all $s \in S$. Thus, S cannot form an unstable cycle. ■

However, note that $\mathbf{x}_\sigma$ may not be Pareto efficient. To see this, consider Example 2 with $\sigma = (\alpha, \beta)$. Here, the allocation $\mathbf{x}_\sigma$ is Pareto dominated by $\mathbf{x}_{\sigma'}$ with $\sigma' = (\beta, \alpha)$. The subset of Pareto efficient serial dictatorship allocations, however, satisfy all of our desiderata; the following lemma shows that these are, in fact, *all of the satisfactory integer allocations*.

Lemma 8 *Suppose that $\mathbf{x}$ is an integer allocation satisfying* **[ER]**, **[QR]**, **[PR]**, **[S]**, *and* **[PE]**. *Then $\mathbf{x} = \mathbf{x}_\sigma$ for some $\sigma \in \Sigma$.*

Proof We argue the claim by induction on q. The base case $q = 1$ is trivial: if c is the category with $q_c = 1$, then any satisfactory integer allocation must give this unit to a highest-priority eligible agent in c, if one exists.

Suppose that the claim holds for all instances with $q = k - 1$, and consider an instance with quota $q = k$. We first show that in any satisfactory integer allocation $\mathbf{x}$ (with $V(\mathbf{x}) > 0$), a highest-priority agent in some category is allocated from that category. Suppose that this were not the case, and consider an agent a who is allocated from category c. By assumption, there is some highest-priority agent a' who is not allocated from c. If a' is unallocated, then $\mathbf{x}$ would violate **[PR]**. Hence, a' must be allocated in some other category c'. By assumption, a' does not have highest priority in c', meaning that the highest-priority agent a'' of c' is not allocated in c'. Continuing this reasoning, we will (by finiteness) eventually revisit an agent and discover an unstable cycle, contradicting that $\mathbf{x}$ satisfies **[S]**.

Now, let c^* be a category allocating to its highest-priority agent, and a^* the highest-priority agent in c^*. We can realize this allocation by having c^* be the first category in the ordering φ. What remains is an allocation problem for agents $\mathcal{A} \setminus \{a^*\}$ to categories $\mathcal{C}$, where the quota of c^* has been

3. For ease of presentation, we ignore ties. This assumption corresponds to each category having a total ordering over eligible agents; in case there are multiple unallocated agents in the same highest priority tier, we can use any fixed tie-breaking rule (alternately, any fixed extension of the total preorder $\succeq_c$).

reduced by 1. Let $\mathbf{y}$ be the restriction of $\mathbf{x}$ to this problem. It is immediate that $\mathbf{y}$ is a satisfactory integer allocation. By our inductive hypothesis, $\mathbf{y}$ can be realized as a serial dictatorship allocation $\mathbf{y}_{\varphi'}$ in this sub-problem. Then, $\mathbf{x}_{(c^*,\varphi')}$ realizes $\mathbf{x}$. ■

Using these lemmas, we can now show that any satisfying (integral) assignment can be realized as the maximizer of (P_δ) for some valid perturbation δ. This shows that our framework allows us to select any desired satisfactory allocation.

Theorem 9 *Let $\mathbf{x}$ be a satisfactory integer allocation. Then, there exists a valid δ such that $\mathbf{x}$ is a solution to (P_δ).*

Proof In the following, it is convenient to argue using *positive* perturbations (i.e., a bonus rather than a penalty). That is, for every $a \in \mathcal{A}, c \in \mathcal{C}$, we set the coefficient of $x_{a,c}$ in the objective as $1 + \rho_{a,c}$, such that $\rho_{a,c} \in [0, \rho_{\max}]$ for all eligible (a, c), and $\rho_{a,c} \geq \rho_{a',c}$ for all $a \succeq_c a'$. To convert the $\rho_{a,c}$ to valid perturbations $\delta_{a,c}$ (Definition 2), we can simply re-scale them by $1 + \rho_{\max}$ to get $\delta_{a,c} = \frac{\rho_{\max} - \rho_{a,c}}{1+\rho_{\max}}$. Then, it is easy to check that these perturbations satisfy Positivity and Consistency. Also, by choosing $\rho_{\max} = \frac{1}{3mn}$, we ensure that $\sum_{a,c} \delta_{a,c} < mn \cdot \rho_{\max}/(1 + \rho_{\max}) < 1/3$; thus, the $\delta_{a,c}$ constitute a valid perturbation.

Let $v := V(\mathbf{x})$. By Lemma 8, $\mathbf{x} = \mathbf{x}_\sigma$ for some ordering $\sigma = (\sigma_1, \ldots, \sigma_q) \in \Sigma$. We may also, without loss of generality, assume that the first v entries of σ result in the allocation of an agent: note that any entry σ_i corresponding to a depleted category can be moved to the end of the ordering without affecting the agents available to any later entry.

Now, we set the perturbations as follows:

1. Let a be the top-ranked agent in the category σ_1. We set $\rho_{a,\sigma_1} = \rho_{\max}$.
2. In stage i, let $r \leq i$ be the lowest ranking of an unallocated agent in category σ_i. Let $r' < r$ be the ranking of the agent most recently allocated in σ_i, with $r' = 0$ if no agent has yet been allocated through σ_i. For $j = r' + 1, r' + 2, \ldots, r$, let a_j be the agent with ranking j in σ_i, and define $A_i = \{a_{r'+1}, a_{r'+2}, \ldots, a_r\}$. We set $\rho_{a_j,\sigma_i} = \rho_{\max}/(n + 1)^{i-1} + (r - j) \cdot \varepsilon$, for some $\varepsilon \ll \rho_{\max}/(n + 1)^n$.

The main invariant maintained by the above construction is that at any stage i, the smallest perturbation $\rho_{a,c}$ for $c = \sigma_i$ and any $a \in A_i$ is greater than the *sum of all perturbations* of (a, c) pairs set in rounds $i' > i$. As a result, the optimal matching among pairs (a, c) considered in rounds i and greater must include at least one pair (a_j, σ_i) for some $a_j \in A_i$. Moreover, since the agents $a_{r'+1}, a_{r'+2}, \ldots, a_{r-1}$ were allocated in rounds prior to i, any optimal matching with respect to the $\rho_{a,c}$ must have $x_{a_r,\sigma_i} = 1$. This exactly corresponds to the outcome x_σ realized via Serial Dictatorship with order σ. Thus, the satisfactory integer allocation x_σ is realized as a solution to (P_δ). ■

4. Perturbed LP Allocations: Insights and Extensions

4.1. Satisfying Additional Desiderata

We first discuss the relation between our perturbed LP solutions and the additional desiderata in Section 2.3. Since these properties compare the allocations of different instances, we will need a procedure for selecting δ that does not depend on specific features of the instance.

Definition 10 (Ranking-Based Allocation Rule) *A ranking-based allocation rule ψ_f is parametrized by a function $f : [n] \to \mathbb{R}$. It computes $\delta_{a,c} = f(r_c(a))$ for all a, c (i.e., the perturbation $\delta_{a,c}$ depends only on the ranking of agent a in c, but not on the identity of a or c or any quotas), and then returns an optimal solution to (P_δ).*

Lemma 11 *Let $f : [n] \to \mathbb{R}$ be a fixed, monotone decreasing function. Then, the ranking-based allocation rule ψ_f satisfies* **[M]**, **[SP]**, *and* **[ST]**.

Proof To prove **[M]**, consider two instances I, I' with the same $\mathcal{A}, \mathcal{C}, \{\succeq_c\}$, and with $q'_c \geq q_c$ for each c. Let $\mathbf{x} \in \psi_f(I)$; that is, $\mathbf{x}$ is an optimal feasible solution to the LP (P_δ) with quotas q_c, where δ is computed from the rankings using f. Enlarging the quotas from q_c to q'_c cannot hurt feasibility, and all perturbation weights are the same for both instances I, I'. If $\mathbf{x}$ remains maximal for I', then $\mathbf{x} \in \psi_f(I')$. Otherwise, we can repeatedly Pareto improve $\mathbf{x}$ by applying the augmentation procedure from the proof of Lemma 6, until the resulting $\mathbf{y}$ is maximal, i.e., an optimal feasible solution for (P_δ) with augmented quotas q'_c. This means that $\mathbf{y} \in \psi_f(I')$, so we have identified a $\mathbf{y} \in \psi_f(I')$ which Pareto dominates $\mathbf{x}$.

For **[SP]**, the way in which the $\delta_{a,c}$ are computed ensures that for any categories c, c' with identical priority orders and any eligible agent $a \succeq_c \theta_c$, $\delta_{a,c} = \delta_{a,c'}$. Therefore, such categories will be treated equivalently in the objective, so they can be merged without affecting the total allocation through these categories.

For **[ST]**, note that by decreasing their priority (i.e., increase their ranking) in category c, agent a will increase the value of $\delta_{a,c} = f(r_c(a))$ (since f is monotone decreasing). This makes it more disadvantageous for c to allocate to a. ■

4.2. Allocation with Transferable Quotas

We next consider a modified setting in which we loosen **[QR]**, and consider alternate desiderata which allow unused quotas to be transferred across categories. In some domains, such as healthcare, the efficiency of an allocation is more important than the strict adherence to quotas; for example, excess medical supplies should not be withheld from patients who do not have certain demographic factors. In these settings, there can be a cost (economic, logistical, moral, etc.) to adjusting the quotas from their pre-determined values. Therefore, a natural goal is to fully allocate the resource in a way that requires minimal deviation from the quotas. Here, a full allocation is one with value

$$v^* := \min\Big\{q, |\{a \in \mathcal{A} : a \succ_c \theta_c \text{ for some } c\}|\Big\},$$

meaning that it either allocates all q resource units, or allocates to every candidate who is eligible in at least one category.

We can measure the *transfer efficiency* of an allocation as the minimum amount of quota that must be transferred between categories to reach a full allocation. Formally, we define the function,

$$T(\mathbf{y}) := \sum_{c \in \mathcal{C}} \Big(\big(\sum_{a \in \mathcal{A}} y_{a,c} \big) - q_c \Big)^+$$

to represent the minimum transfer amount. Then, the following desideratum captures our notion of transfer efficiency.

[TE] Transfer Efficient: Among all full allocations, $\mathbf{x}$ minimizes the total amount of transferred quota: $\mathbf{x} \in \operatorname{argmin}_{\mathbf{y}} \Big\{ T(\mathbf{y}) \Big\}$.

The following lemma describes the close relationship between our two notions of efficiency.

Lemma 12 *Suppose that* $\mathbf{x}$ *is an allocation satisfying* **[TE]** *and* $\hat{\mathbf{x}}$ *is an allocation satisfying* **[PE]**. *Then,* $T(\mathbf{x}) = v^* - V(\hat{\mathbf{x}})$.

Proof First, we argue that $T(\mathbf{x}) \leq v^* - V(\hat{\mathbf{x}})$. By total unimodularity, we may assume that $\hat{\mathbf{x}}$ is integral. Consider the procedure given in Algorithm 2 to transfer some quotas in $\hat{\mathbf{x}}$.

Algorithm 2 Transfer Quota

1: **while** there is an unallocated agent eligible in a category c and a category c' with unused quota **do**
2: let a be the highest-priority unallocated agent in c.
3: transfer one unit of quota from c' to c and allocate a through c.

When this procedure terminates with an allocation $\hat{\mathbf{y}}$, either every agent eligible in a category has been allocated, or all quotas have been filled. Thus, $V(\hat{\mathbf{y}}) = v^*$. The Pareto efficiency of $\hat{\mathbf{x}}$ ensures that every category with an unallocated agent has met its quota. Therefore, every unit transferred to a category c will be in excess of its quota. Additionally, no category that receives additional quota will donate it later in the procedure: all received quota is immediately allocated. Thus, the total amount of transferred quota is exactly equal to the increase in allocation amount

$$V(\hat{\mathbf{y}}) - V(\hat{\mathbf{x}}) = v^* - V(\hat{\mathbf{x}}) = \sum_{c \in \mathcal{C}} \Big(\big(\sum_{a \in \mathcal{A}} y_{a,c} \big) - q_c \Big)^+.$$

Note that $\hat{\mathbf{y}}$ was one allocation in the minimization defining $\mathbf{x}$, so $T(\mathbf{x}) \leq v^* - V(\hat{\mathbf{x}})$.

Next, we argue that $T(\mathbf{x}) \geq v^* - V(\hat{\mathbf{x}})$. Consider the flow network from Fig. 3. By increasing the capacity of one (c, t) edge in this network by Δ, the value of the maximum flow (i.e. the total allocation) increases by at most Δ. Similarly, by decreasing the capacity on a (c, t) edge, we do not increase the value of the maximum flow. Thus, the transfer of Δ units of quota results in an increase of at most Δ in the total allocation. By definition, the maximum total allocation subject to the quota constraints is $V(\hat{\mathbf{x}})$, so we must transfer at least $v^* - V(\hat{\mathbf{x}})$ units of quota to realize an allocation of value v^*. Hence, $T(\mathbf{x}) \geq v^* - V(\hat{\mathbf{x}})$. ■

Thus, we can use our perturbed LP (P_δ) from Section 3 to find a Pareto efficient solution $\hat{\mathbf{x}}$, and then use the procedure in the proof to modify $\hat{\mathbf{x}}$ to a transfer efficient allocation $\hat{\mathbf{y}}$. In fact, $\hat{\mathbf{y}}$ additionally satisfies **[ER]**, **[PR]**, and **[S]**. The first is immediate from the definition, and the latter two follow by arguments analogous to Theorem 3.

4.3. Unanimous Agents

We define *unanimous* agents to be those who are allocated by every integer allocation satisfying **[ER]**, **[QR]**, **[PR]**, **[PE]**, and **[S]**. The unanimous agents are crucial to any allocation procedure targeting these desiderata, as such a procedure must decide through which categories each of these

agents is allocated. There is an equivalent characterization of unanimous agents that allows them to be identified in polynomial time.

For a given agent a and a category c for which a is eligible, we define the *a-restriction* of c to be the total preorder $\succ_{c\setminus a}$ obtained from $\succ_c$ by moving a from ranking $r_c(a)$ to ranking $r_c(a)+1$, and placing θ_c immediately above a (thereby making a, and all agents a' with lower priority than a in c, ineligible for c).

Lemma 13 *Let V^* be the value of (P) on the original instance. For a given agent a, let $V^*_{\setminus a}$ be the value of (P) on the instance where $\succ_c$ has been replaced with the restriction $\succ_{c\setminus a}$ in each category c for which a is eligible. Then, a is unanimous if and only if $V^* > V^*_{\setminus a}$.*

Proof We argue the forward direction by its contrapositive. Suppose that $V^* = V^*_{\setminus a}$. Then, we have found an integer allocation with value V^* satisfying **[ER]**, **[QR]**, and **[PE]** on the a-restricted instance. By Corollary 4, there must also be an integer allocation with value V^* that additionally satisfies **[PR]** and **[S]**. Note that a is not present on any of the a-restricted lists, so is not allocated. However, this allocation is also feasible for the original priority lists. Since we have located a satisfactory integer allocation that does not include a, a is not unanimous.

We also argue the reverse direction by its contrapositive. Suppose that a is not unanimous. Then, there is a satisfactory integer allocation in which a is not allocated. By definition, this allocation has value V^*. By Axiom 3, no category could allocate to an agent with lower priority than a. Thus, this allocation is also feasible for the a-restricted instance, so $V^* = V^*_{\setminus a}$. ■

As two immediate corollaries to this lemma, we can derive two sufficient conditions for an agent to be unanimous.

Corollary 14 *Let V^* be the value of (P). Then, agent a is unanimous if the union of all eligible agents in the a-restricted instance has cardinality less than V^*.*

Corollary 15 *Let ℓ_c be the number of eligible agents in category c. Then, an agent a is unanimous if they are in the top $\min\{\ell_c, q_c\}$ agents in c.*

5. Selecting Equitable Allocations via Ranking-Based Perturbations

Thus far, we have addressed existence, efficiency and computation of satisfactory allocations. In particular, Theorem 3 tells us that any choice of valid δ induces a satisfactory allocation, and Theorem 9 shows that every satisfactory allocation is induced by some δ.

In this section, we take a more principled approach to the question of allocation selection. We discuss how our perturbed LP approach gives us the freedom to set δ to select *particular* allocations satisfying additional notions of equity. We restrict our attention to ranking-based allocation rules, so that their allocations additionally satisfy **[M]**, **[SP]**, and **[ST]** by Lemma 11.

5.1. Minimizing the Average Ranking

The ranking of an agent a in category c provides a proxy for how important it is to c that a be allocated. Thus, a natural notion of the quality of an allocation is the average rank of the allocated items in their category — if this average rank is small, it means that categories predominantly allocate their units to their top choices. This notion of average ranking has been previously considered in the context of Stable Marriage (Pittel, 1989).

Lemma 16 *Consider the ranking-based allocation rule ψ_f with $f(k) = \frac{k}{4mn^2}$. The allocations induced by ψ_f are satisfactory and minimize the average ranking of the allocated agents among all satisfactory allocations.*

In other words, to minimize the average rank, we can choose perturbation penalties that grow *arithmetically* in the agent ranking.

Proof Fix rankings, and define δ^A as $\delta^A_{a,c} = f(r_c(a)) = \frac{r_c(a)}{4mn^2}$. For the first claim, it suffices to argue that δ^A is valid by Theorem 3. By construction, each $\delta^A_{a,c}$ is positive. Since each $r_c(a) \leq n$,

$$\sum_{a\in\mathcal{A}}\sum_{c\in\mathcal{C}} \delta^A_{a,c} \leq \sum_{a\in\mathcal{A}}\sum_{c\in\mathcal{C}} \tfrac{1}{4mn} = \tfrac{1}{4} < \tfrac{1}{3},$$

so δ^A has small effect. δ^A is also consistent, because $r_c(a) \leq r_c(a')$ if and only if $a \succeq_c a'$.

For the second claim, we consider the objective value of (P_{δ^A}). We have

$$V_{\delta^A}(\mathbf{x}) = V(\mathbf{x}) - \sum_{a\in\mathcal{A}}\sum_{c\in\mathcal{C}} \delta^A_{a,c}\cdot x_{a,c} = V(\mathbf{x}) - \tfrac{1}{4mn}\cdot\Big(\tfrac{1}{n}\sum_{a\in\mathcal{A}}\sum_{c\in\mathcal{C}} r_c(a)\cdot x_{a,c}\Big).$$

$V(\mathbf{x})$ is be the same for all satisfactory (and thus all Pareto efficient) allocations. The parenthesized expression is exactly the average ranking of all allocated agents. Thus, solutions to (P_{δ^A}) minimize this average ranking among all satisfactory allocations. ■

5.2. Minimizing the Maximum Ranking

While the average rank is a natural utilitarian notion of allocation quality, another natural alternative that is more equitable is to try and minimize the maximum allocated ranking across categories. The maximum ensures a stricter notion of "fairness," in that it discourages even one category using its quota to allocate to an agent lower in its ranking. Notice that for the maximum ranking, it is more meaningful to focus solely on integer allocations, to avoid a discontinuity as an allocation goes to 0. In addition, minimizing the maximum ranking makes most sense as a notion of equity in settings where all of the quotas are roughly equal.

Lemma 17 *Consider the ranking-based allocation rule ψ_f with $f(k) = \frac{1}{4mn}\cdot\left(\frac{1}{n+1}\right)^{n-k}$. The integer allocations induced by ψ_f are satisfactory and minimize the maximum ranking of the allocated agents among all satisfactory integer allocations.*

In other words, to minimize the maximum rank, we can choose perturbation penalties that grow *geometrically* in the agent ranking.

Proof Fix rankings, and define δ^G as $\delta^G_{a,c} = f(r_c(a)) = \frac{1}{4mn}\cdot\left(\frac{1}{n+1}\right)^{n-r_c(a)}$. For the first claim, it suffices to argue that δ^G is valid by Theorem 3. By construction, each $\delta^G_{a,c}$ is positive. Since each $r_c(a) \leq n$, we have that $\left(\frac{1}{n+1}\right)^{n-r_c(a)} \leq 1$. Thus,

$$\sum_{a\in\mathcal{A}}\sum_{c\in\mathcal{C}} \delta^G_{a,c} \leq \sum_{a\in\mathcal{A}}\sum_{c\in\mathcal{C}} \tfrac{1}{4mn} = \tfrac{1}{4} < \tfrac{1}{3},$$

so δ^G has small effect. δ^G is also consistent, because $r_c(a) \leq r_c(a')$ if and only if $a \succeq_c a'$, and $\delta^G_{a,c}$ is an increasing function in $r_c(a)$. For the second claim, we consider the objective value of (P_{δ^G}):

$$V_{\delta^G}(\mathbf{x}) = V(\mathbf{x}) - \sum_{a \in \mathcal{A}} \sum_{c \in \mathcal{C}} \delta^G_{a,c} \cdot x_{a,c}.$$

Restricting our attention to integer allocations $\mathbf{x}$, the sum can be rewritten as

$$\frac{1}{4mn \cdot (n+1)^n} \cdot \sum_{(a,c):x_{a,c}=1} (n+1)^{r_c(a)}. \tag{2}$$

Let $R(\mathbf{x}) := \max_{(a,c):x_{a,c}=1} \{r_c(a)\}$. By definition, for a fixed $R(\mathbf{x})$, the value of the sum in Eq. (2) falls in the interval $\left[(n+1)^{R(\mathbf{x})}, n \cdot (n+1)^{R(\mathbf{x})}\right]$. Since these intervals are non-overlapping, the integral allocation maximizing V_{δ^G}, so minimizing this sum, must also minimize $R(\mathbf{x})$. ■

6. Conclusions

The problem of reserve allocations is central in many real-world settings, as many allocation criteria can be expressed through priorities and quotas. In this work, we built on a model of Pathak et al. (2021) that modelled reserve allocation as a bipartite matching problem with additional hard priority constraints. We then showed that a *valid* allocation – those which obey priorities and quotas, and also are Pareto efficient – can be located using a simple b-matching LP. In more detail, we perturb edge weights in a way that enforces the priorities, but also leverages the integrality of the corner points of the b-matching polytope to only select valid allocations. Moreover, by introducing a stability criterion from the perspective of the categories, we were able to give a complete algorithmic characterization of all valid allocations.

The clean formulation of our algorithm has many benefits. First, it amounts to the computation of a single weighted b-matching, making it more efficient than previous algorithms in this space. Beyond this, we were able to utilize the LP structure to establish many additional properties of our allocations. Finally, the under-specification of our algorithm's parameters provided an opportunity to secondarily enforce notions of fairness. Determining other secondary objectives to which this framework is amenable is an interesting direction for future work. For example, the ability to incorporate notions of utility — both on the part of the categories and the agents — could relax our assumption on indifference of allocation and provide a more realistic setting for settings with economic incentives.

While our flexible approach can locate a large collection of potential allocations, in some settings, it may be desirable to specify a natural rule (equivalently, add additional desiderata) that, given any priorities and quotas, identifies a unique "best" allocation. While the work of Delacrétaz (2021) makes an effort in this direction, the additional axiom proposed there is arguably not fully natural, but more importantly, can result in unexpected or inefficient allocations. An ambitious goal would be to locate an allocation that (1) is the unique allocation meeting a set of criteria, (2) can be realized as the maximizer of some objective function over allocations, and (3) can be located by an efficient algorithm. We hope the techniques introduced here can help in building towards this goal.

References

Atila Abdulkadiroğlu, Parag A Pathak, Alvin E Roth, and Tayfun Sönmez. The Boston public school match. *American Economic Review*, 95(2):368–371, 2005.

Erin E Andrews, Kara B Ayers, Kathleen S Brown, Dana S Dunn, and Carrie R Pilarski. No body is expendable: Medical rationing and disability justice during the Covid-19 pandemic. *American Psychologist*, 76(3):451, 2021.

Orhan Aygun and Bertan Turhan. How to de-reserve reserves. *Available at SSRN 3801466*, 2021.

Haris Aziz and Florian Brandl. Efficient, fair, and incentive-compatible healthcare rationing. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 103–104, 2021.

Nawal Benabbou, Mithun Chakraborty, Xuan-Vinh Ho, Jakub Sliwinski, and Yair Zick. Diversity constraints in public housing allocation. In *17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018)*, 2018.

Nawal Benabbou, Mithun Chakraborty, and Yair Zick. Fairness and diversity in public resource allocation problems. *Bulletin of the Technical Committee on Data Engineering*, 2019.

Charles E Binkley and David S Kemp. Ethical rationing of personal protective equipment to minimize moral residue during the Covid-19 pandemic. *Journal of the American College of Surgeons*, 230(6):1111–1113, 2020.

Péter Biró and Jens Gudmundsson. Complexity of finding Pareto-efficient allocations of highest welfare. *European Journal of Operational Research*, 291(2):614–628, 2021.

Jose Correa, Natalie Epstein, Rafael Epstein, Juan Escobar, Ignacio Rios, Nicolás Aramayo, Bastian Bahamondes, Carlos Bonet, Martin Castillo, Andres Cristi, et al. School choice in Chile. *Operations Research*, 2021.

COVAX. Covax explained. https://www.gavi.org/vaccineswork/covax-explained, 2020. Accessed: 2022-02-14.

David Delacrétaz. Processing reserves simultaneously. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 345–346, 2021.

Ezekiel J Emanuel, Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P Phillips. Fair allocation of scarce medical resources in the time of Covid-19, 2020.

Aytek Erdil and Haluk Ergin. Two-sided matching with indifferences. *Journal of Economic Theory*, 171:268–292, 2017.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Parag A Pathak, Tayfun Sönmez, M Utku Ünver, and M Bumin Yenmez. Fair allocation of vaccines, ventilators and antiviral treatments: leaving no ethical value behind in health care rationing. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 785–786, 2021.

Boris Pittel. The average number of stable matchings. *SIAM Journal on Discrete Mathematics*, 2 (4):530–549, 1989.

Lyle Ramshaw and Robert E Tarjan. On minimum-cost assignments in unbalanced bipartite graphs. *HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2012-40R1*, 2012.

Daniela Saban and Jay Sethuraman. The complexity of computing the random priority allocation matrix. *Mathematics of Operations Research*, 40(4):1005–1014, 2015.

Tayfun Sönmez, M Bumin Yenmez, et al. *Affirmative action in India via vertical and horizontal reservations*. Boston College, 2019.

Douglas B White and Bernard Lo. A framework for rationing ventilators and critical care beds during the Covid-19 pandemic. *Journal of the American Medical Association*, 323(18):1773–1774, 2020.

Appendix A. Existing Allocation Rules (and their Shortcomings)

Here, we summarize some allocation techniques from the prior literature.

In Section 3 of their work, Pathak et al. (2021) consider a special case of our model where all agents are eligible for all categories and where the priority order of each category is total. They define the collection of *sequential reserve matchings* $\{\varphi_{\triangleright}\}$ as those computed by Algorithm 3 for each total ordering $\triangleright$ over the categories.

Algorithm 3 Sequential Reserve Allocation

```
for each category c in order of ▷ do
    while c has remaining quota and an (eligible) unallocated agent a do
        allocate one unit to a through c
```

It is immediate from this description that every sequential reserve matching $\varphi_{\triangleright}$ satisfies **[ER]**, **[PR]**, and **[QR]**. In addition, sequential reserve matchings, and indeed, any form of serial dictatorship, also satisfy **[S]**. However, while **[PE]** is trivially satisfied in their restricted setting, it is not guaranteed when some agents are ineligible for some categories, as illustrated by Example 2.

Example 2 *([PE] violations by Algorithm 3) Consider the following instance.*

α (1)	β (1)
a	a
b	

If $\alpha \triangleright \beta$, then the sequential allocation $\varphi_{\triangleright}$ has α allocate to a and leaves b unallocated. This is Pareto dominated by the allocation to b through α and to a through β.

In follow-up work, Delacrétaz notes another inherent unfairness of sequential allocation; categories that are processed later will have more opportunity to allocate to agents lower in their priority lists (Delacrétaz, 2021). To address this, he introduces a category neutrality desideratum wherein an agent only receives a greater allocation from eligible category c than c' if c' has allocated its full quota to higher-priority agents. Delacrétaz describes a 'water-filling' procedure called *Simultaneous Allocation* that produces a category neutral allocation.

We refer readers to (Delacrétaz, 2021) for the formal definition of the Simultaneous Allocation rule. The most notable feature of the rule is that each agent a receives an equal allocation[4] from every category that a qualifies for. Unfortunately, as Delacrétaz notes, this procedure cannot ensure Pareto efficiency under partial eligibility lists; for example, in the instance in Example 2, under Simultaneous Allocation, a is allocated $\frac{1}{2}$ unit each from categories α and β, for a total allocation of 1, while b is allocated the remaining $\frac{1}{2}$ unit. This fractional allocation is again Pareto dominated by the integer allocation giving a one unit through category β and b one unit through category α. In addition, the simultaneous allocation procedure also fails to be stable, even in the case of total priority orders.

4. subject to quota and priority constraints.

Example 3 *(Simultaneous allocation is not stable) Consider the following instance.*

α (1)	β (1)	γ (1)
a	b	c
b	c	a
c	a	b

The simultaneous allocation procedure gives each agent $\frac{1}{3}$ of a unit through each category. This allocation has an unstable cycle $(c,\alpha) \to (a,\beta) \to (b,\gamma) \to (c,\alpha)$. On the other hand, the allocation where each category gives a full unit to its highest-priority agent is the unique stable and Pareto-efficient allocation.

Moreover, the equal allocation property also leads to violation of Sybil-proofness, and indeed, merging/splitting categories can result in Pareto improvements

Example 4 *(Merging categories can increase allocation) Consider the following instance:*

α (1)	β (1)	γ (1)
a , b	a , b	a , b
	c	c

The simultaneous allocation procedure will require β and γ to each allocate $\frac{1}{3}$ of a unit to each of a and b, leaving a total of $\frac{2}{3}$ to allocate to c (and giving their eligible agents a total allocation of $\frac{8}{3}$). However, if β and γ merge, simultaneous allocation will require the merged category to contribute half a unit to each of a and b, allowing them to contribute a full unit to c (and giving their eligible agents a total allocation of 3).

Lastly, we consider the reverse rejecting rule of Aziz and Brandl (2021), which iteratively removes agents from consideration as long as there remains a maximal allocation.

Algorithm 4 Reverse Rejecting Allocation

$R \leftarrow \varnothing$, $M^* \leftarrow$ maximal allocation (without **[PR]** considerations)
for each agent a in an (arbitrary) order $\succ_\pi$ **do**
  $M \leftarrow$ maximal allocation in instance where $R \cup \{a\}$ and all lower priority agents are deleted from each category
  if $|M| = |M^*|$ **then**
    $R \leftarrow R \cup \{a\}$, $M^* \leftarrow M$
return M^*

This algorithm turns out to satisfy most of our primary desiderata (**[ER]**, **[QR]**, **[PR]**, and indirectly **[PE]**); however, it does not ensure stability. In particular, the algorithm underspecifies the selection of the matching M, so it permits the selection of an unstable matching, for example $\{(c,\alpha),(a,\beta),(b,\gamma)\}$ in Example 3 (and since the choice of the final matching M^* is agnostic of all priority information, it is unclear if this can fixed).