Social Choice and Welfare
https://doi.org/10.1007/s00355-025-01589-8

ORIGINAL PAPER

Quota implementation in assignment games

Xu Lang[1,2] · **Jiahui Li**[3,4]

Received: 26 March 2024 / Accepted: 24 February 2025

Abstract
We study the implementability of quota policies in assignment games with constraints, where the planner can set a quota system before the allocation of resources. The distributional constraint specifies demand floors for subgroups of agents. In addition, the planner contends with a group influence constraint, where a larger group of agents exercises more significant influence on a quota policy. We completely characterize the set of all implementable quota systems and provide a game-theoretic interpretation for the implementability condition. We also characterize the extreme points of the polytope defined by the implementable demand floor quotas, and study optimal demand floors in a class of two-stage quota-setting games.

1 Introduction

Many assignment and matching markets are regulated by various forms of distributional constraints, e.g., explicit quotas and non-quota schemes like race-conscious policies. One prominent example is the “regional cap" in the Japanese residency matching program that matches hospitals with doctors (Kamada and Kojima 2015). To regulate the geographical distribution of doctors, the total number of doctors matched within a region is subject to a “regional cap." Another example is the majority quotas in the Boston school choice program, where diversity constraints limit the demographic distributions of admitted students at public schools (Abdulkadiroğlu and Sönmez

We thank the Advisory Editor and two anonymous referees for their comments.

✉ Xu Lang
langxu1011@gmail.com

Jiahui Li
jiahui_li@zufe.edu.cn

1 Center for Economic Research, Shandong University, Jinan, China

2 Center for Research on Experimental and Theoretical Economics, Shandong University, Jinan, China

3 School of Economics, Zhejiang University of Finance and Economics, Hangzhou, China

4 Center for Economic Behavior and Decision-making (CEBD), Neuro and Behavior EconLab (NBEL), Zhejiang University of Finance and Economics, Hangzhou, China

Published online: 26 March 2025

2003).[1] Similarly, in the Chinese college admission problem (*Gaokao*), every university has its own quota system on the numbers of admitted students from different provinces as well as out-of-province quotas for minority groups. The quota policies are also widely used in other settings. During the Covid-19 pandemic, the policymakers and healthcare professionals implemented reserve policies to ensure the equitable distribution of ventilators and intensive care units (ICUs) for various regions and population groups (Pathak et al. 2023). Such quota policies are highly controversial and have continuously influenced ballot initiatives, lawsuits, and public opinion.[2]

In many markets, quotas remain stable over long periods and are often taken as given in the economic analysis. Anecdotally, these quotas are shaped by norms, customs, and precedents, as well as legislative bargaining among interested parties over years and decades. For example, the use of explicit quota systems for minority applicants in the U.S. university admissions was ruled down in Bakke v. California Board of Regents, and Gratz v. Bollinger respectively, while Grutter v. Bollinger permitted race-conscious policies. Another example is the admission quota system of Tsinghua university in China, which dates back to the time when the school was initiated, where the admission quotas for each province were distributed according to the financial contribution of each province. Since any change in the existing quota system may have a strong redistributive effect on many interested parties, reforms of quota systems are usually either politically infeasible or less progressive.[3]

However, quotas are changing regularly over a certain period of time in some markets of practical interest. The policymakers can adjust quotas for under-represented groups based on historical data and demographic trends. For example, in school choice, if a certain ethnic group is under-represented compared to the regional population, the quota for that group can be increased to achieve a more balanced representation.[4] During the Covid pandemic, quota policies in many countries underwent periodic reviews, taking into account the changing dynamics of the pandemic. Adjustments were made based on shifts in infection rates and the availability of testing supplies. Often, quota policies are influenced not just by policymakers but also by the lobbying efforts of various interested groups. Economies of scale often occur when groups unite their forces, e.g., combining two groups has a greater influence on a quota policy than if the groups act independently (Shapley 1971; Austen-Smith and Wright 1994).[5]

[1] In June 2023, overturning decades of precedent, the U.S. Supreme Court ruled that it is unconstitutional for colleges, universities, and professional schools to consider race as one factor in deciding whom they will admit.

[2] For the literature on race-conscious affirmative actions, see for example Chan and Eyster (2003), Fryer et al. (2008), and Chan and Eyster (2009), where the literature finds that color-blind affirmative action policy is less efficient than the optimal color-sighted policy that achieves the same degree of racial diversity.

[3] An example is Proposition 209 in California (1996), which banned affirmative actions in public employment, education, and contracting. Critics argue that this reform led to a decrease in diversity at California's public universities. This reform has been seen by some as a step backward in promoting equal opportunities.

[4] The evolution of quotas in college admissions in China is an example. Each year, when the central government pushes local authorities to draft their plans before the end of the year, it will clarify how to adjust admission quotas for different regions, which is the crucial factor for local authorities to adjust their policies. The final outcome of each school's quotas is determined by the bargaining among three parties: the central government, local authorities, and the school.

[5] Consider a doctor-hospital matching problem. If hospitals from both regions form a coalition, their combined demand might carry more weight. They could argue that boosting physician numbers in both

In this paper, we study the implementability of quota policies in two-sided assignment models (e.g., Shapley and Shubik 1971; Bogomolnaia and Moulin 2001; Budish et al. 2013). We assume that the planner can set a collection of distributional constraints before running a market, including demand floor quotas for subgroups of agents and supply ceiling quotas for subsets of goods. In addition, the planner faces a class of lobbying constraints which we call group influence constraints. For the demand side, we assume a combination of two groups of agents has a greater influence on the quota policy than two separate groups. For the supply side, we assume a combination of two groups of suppliers has a smaller influence than two separate groups (e.g., due to public interests). We formally introduce such group influence constraints by assuming demand floor quotas to be supermodular and supply ceiling quotas to be submodular. We investigate a joint implementation of the distributional constraints and group influence constraints.

Using a network flow approach, we completely characterize all implementable quota systems. Intuitively, the implementability condition requires that for any subset of agents and any subset of goods, the demand and supply are approximately balanced. We relate this condition to convex games (e.g., Shapley 1971), where we establish an equivalence between the implementability and the non-emptyness of the 'joint core' in a pair of games induced by quotas. We then use the implementability condition to study a two-stage quota design game, where at the first stage, the planner sets demand floor quotas with ceiling quotas being given, and in the second stage, a market works in a decentralized way given the quotas. We provide a characterization of the set of extreme points of demand floor quotas. Extremal quotas are useful: When the objective is linear or convex in quotas, the optimal solution is guaranteed to occur at an extreme point; When the objective is nonlinear, the extreme points provide a finite set of candidate solutions and reduce the search space to a manageable set.

Our implementation problem and results are closely related to Budish et al. (2013) in several aspects. Budish et al. (2013) characterize the set of feasible random allocations for indivisible goods and obtain a generalization of the Birkhoff-von Neumann theorem for bihierarchical constraints. In their implementation problem, they ask whether a random allocation can be decomposed as a convex combination of feasible pure allocations. Their problem assumes that quotas are fixed and implementable, while in our problem, quotas are variable. Hence, the two implementation problems are solved at two different stages and are complementary for the design and implementation. Different from their universal implementation, we consider a quota-dependent implementation, where the group influence constraints are present as a side-constraint on quotas.

Our approach to optimal quota design borrows tools from the reduced-form auctions (e.g., Matthews 1984; Border 1991; Che et al. 2013; Goeree and Kushnir 2023; Lang and Mishra 2024) and network flow approach to mechanism design (e.g., Vohra 2011; Che et al. 2013). Similar to characterizing the implementability condition in reduced-form auctions, characterization of implementable quotas is required to find an optimal quota system. Che et al. (2013) develop a network flow approach to obtain a reduced-

areas would improve overall health outcomes. Moreover, lobbying together could underscore the broad nature of doctor shortages, making their case more compelling to policymakers.

form characterization with paramodular constraints on quotas. We use the network flow approach for our quota-dependent characterization. In contrast to their implementation on multiunit auction markets, which is one-sided, we consider an assignment market and our notion of sub(super-)modularity is two-sided.

Literature Review. Our paper contributes to the literature on market design with reservation policies. Kojima (2012) shows that majority quotas on the number of majority students can hurt minority students. To overcome the detrimental effect of majority quotas, Hafalir et al. (2013) introduce policies of minority reserves. Ehlers et al. (2014) and Fragiadakis and Troyan (2016) discuss minority reserves with multiple priority levels and mechanisms with hard lower quotas. Echenique and Yenmez (2015) introduce choice functions that reflect the diversity constraints but also satisfy the substitutes property. Nguyen and Vohra (2019) study stable matching with proportionality constraints that require proportional soft bounds instead of ex ante absolute numbers of quotas. Celebi and Flynn (2021) study the trade-offs between using minority quotas and score subsidies in affirmative action. Pathak et al. (2023) propose a reserve system for medical resources with multiple categories and a category-specific priority order is used to prioritize individuals for units in each category. Our paper differs from the existing literature in several aspects: The literature usually assumes that quotas are fixed (e.g., a hard constraint) while we assume them to be a policy choice. Second, we assume a quota designer itself may be subject to side-constraints on quotas. Finally, the literature focuses on design of algorithms that find desirable (e.g., stable or fair) outcomes given reserves and rarely optimizes a global objective. Instead, we assume that the planner has a welfare objective and can optimize over all implementable quotas.

There is a recent small literature on studying quotas as a design variable in matching problems. Afacan et al. (2024) study an allocation of extra quotas above some baseline quotas in school choice. They introduce a constrained efficient matching which is fair and efficient at some (implementable) quota and is not Pareto dominated by other fair and efficient quotas at other quotas. They introduce a simple myopic algorithm that finds constrained efficient matchings among all quotas. Kumano and Kurino (2024) introduce an ex-post student-optimal stable matching that is stable at some (implementable) quota, and is not Pareto dominated by any stable matching at other quotas, and define the associated quota as optimal. They propose a quota adjustment process that finds the optimal quota. One difference between their papers and ours is that in their models the optimal quotas and matchings are simultaneously found by algorithms, while we assume quotas and market outcome are determined in two independent steps. Bobbio et al. (2024) study the problem of jointly deciding how to allocate a budget of additional quotas and finding a student-optimal assignment by an integer programming approach. While their model does not consider distributional constraints, their program can further incorporate such constraints and the implementability results can become useful.

2 Model

Let N be a finite set of n agents (e.g., students, hospitals, patients, buyers) and let O be a finite index set of m goods (e.g., schools, doctors, medical resources, sellers).

Each type $j \in O$ of goods can have more than one units; that is, it can be either private goods or club goods. Each agent can demand more than one type of goods. Let $E = N \times O$ denote the set of all possible agent-good pairs. An **assignment** is described as a matrix $x = (x(i, j)) \in \mathbb{R}_+^E$ indexed by all agents and goods, where each entry $x(i, j)$ is a real-valued quantity of good j assigned to agent i. Without loss of generality, we assume that agent i demands at most one unit of good j:

$$0 \leq x(i, j) \leq 1, \quad \forall\ (i, j) \in E. \tag{1}$$

For any vector $x \in \mathbb{R}^E$, $I \subseteq N$, and $J \subseteq O$, we denote $x(I \times J)$ as the sum over $i \in I$ and $j \in J$ of $x(i, j)$.

Distributional constraints. Distributional policies can be imposed on agent-good pairs. For illustration, we focus on demand floor and supply ceiling constraints. For every subset $A \subseteq N$ of agents, let $d(A) \in \mathbb{R}_+$ denote the floor quotas assigned to A for all goods, with $d(\emptyset) = 0$. The demand floor constraints are represented by

$$x(A \times O) \geq d(A), \ \forall\ A \subseteq N. \tag{2}$$

For each subset $B \subseteq O$ of goods, let $c(B) \in \mathbb{R}_+$ denote the ceiling quotas assigned to B for all agents, with $c(\emptyset) = 0$. The supply ceiling constraints are given by

$$x(N \times B) \leq c(B), \ \forall\ B \subseteq O. \tag{3}$$

We call (c, d) a **quota system**. We say an assignment $x = (x(i, j))$ is **feasible** if x satisfies (1), (2), and (3). Notice that we assume divisible goods which only require (c, d) to be nonnegative and real-valued. So we will not distinguish integral and fractional assignments.

Our model covers many important practical situations. The examples of the demand floor constraints include the minimum numbers of course seats for the students from each department; the minimum quotas for rural hospitals in the matching programs; and the reserves of medical resources for the groups with different traits. The examples of the supply ceiling constraints cover school choice with the majority quotas and auctions where the large suppliers are regulated by antitrust policies.

Group influence constraints. We assume the planner must deal with power constraints on quotas besides the distributional constraint. For the demand side, consider two disjoint groups of agents that can form a large coalition to require a larger demand floor quota. The coalition can convey broader needs and has a higher influence on the policy than the two smaller groups. For the supply side, suppose two groups of firms form a coalition to advocate a larger quota. The regulator, who concerns unfair competition, may trigger more intensive regulation.[6] This could lead to a reduced influence on policy compared to those of independent firms.

[6] Some literature discussed set-asides and competition policies in auctions. Pai and Vohra (2012) show that a flat subsidy can be the most efficient auction design that achieves a distributional requirement. Athey et al. (2013) find that compared to set-asides, subsidizing small bidders would increase revenue with little efficiency cost.

To model these group influence constraints, we introduce the following notions of complementarity and substitutability. Let U be a finite ground set. We say a set-function $f : 2^U \to \mathbb{R}_+ \cup \{-\infty\}$ is supermodular if (i) $f(\emptyset) = 0$ and (ii) f satisfies

$$f(S \cap T) + f(S \cup T) \geq f(S) + f(T), \ \forall \ S, T \subseteq U.$$

We define f as submodular if $-f$ is supermodular and as modular if (ii) holds with equality.

Definition 1 (c, d) satisfies **two-sided paramodular** group influence constraints, if $c : 2^O \to \mathbb{R}_+$ is submodular and $d : 2^N \to \mathbb{R}_+$ is supermodular. We say (c, d) is **modular**, if both c and d are modular.

Supermodular demand floor quotas require that for any two disjoint coalitions of agents, S and T, their joint influence on the quota policy (e.g., $d(S \cup T)$) is greater or equal to their independent influences (e.g., $d(S)$ and $d(T)$ respectively). Submodular supply ceiling quotas reflect a scenario where the government's public initiatives (e.g., competition and health policies) are the priorities such that interested groups (e.g., firms, hospitals, etc.) opposed to the initiatives have less room to influence the quota policy: (1) The regulator is more skeptical of a very large lobbying group and imposes a more stricter regulation; (2) A larger lobbying group has a higher coordination cost due to regulation. Modular problems assume that there is no complementarity or substitutability.

Remark 1 Che et al. (2013) introduce paramodular constraints in multi-unit auctions. Their model requires sub(super-)modularity on the same ground set defined by the agent set, and hence is a problem with one-sided market. Our model assumes sub(super-)modularity on two different ground sets, which reflect the two-sided nature of assignment markets.

We are ready to introduce the notion of implementable quota systems for a planner.

Definition 2 A quota system (c, d) is **implementable**, if there exists a feasible assignment $x \in \mathbb{R}^E$ under (c, d).

2.1 Coalitional games induced by quotas

Before presenting our main result, we discuss an alternative approach to our model from a cooperative game perspective.[7] A given quota system can be decomposed into two TU games: a value (sharing) game and a cost (sharing) game, where quantities are now measured by monetary units. In a value game, a demand quota for a coalition represents the secured payoff of the coalition, while in a cost game, a supply quota represents the secured cost of a coalition. Different from classical problems, the two games are not played independently: for each pair (i, j), the payoff of player $i \in N$ is the cost of player $j \in O$. This implies the stable allocations of the two games must be coupled, e.g., a stable allocation in one game may not be implementable without joining of the players in the other game, and vice versa.

[7] We thank the Advisory Editor for suggesting us this issue.

Formally, consider a quota system (c, d) over N and O. Define a value game (N, d) with player set N and characteristic function d. Define a cost game (O, c) with player set O and cost function c. We call $((N, d), (O, c))$ a pair of **marginal games**.[8] For any value distribution $x \in \mathbb{R}^N$, let $x(A) := \sum_{i \in A} x_i$ for each $A \subseteq N$. For any cost distribution $x \in \mathbb{R}^O$, let $x(B) := \sum_{j \in B} x_j$ for each $B \subseteq O$. Then for each marginal game, **the core** of the game is defined by

$$core(N, d) = \{x \in \mathbb{R}^N | x(N) = d(N),\ x(A) \geq d(A),\ \forall A \subseteq N\},$$

and

$$core(O, c) = \{x \in \mathbb{R}^O | x(O) = c(O),\ x(B) \leq c(B),\ \forall B \subseteq O\}.$$

Suppose further $c(N) = d(O)$. In this case, a feasible assignment x must satisfy

$$x(N \times O) = d(N) = c(O). \tag{4}$$

For a feasible assignment $x \in \mathbb{R}^E$, define the **marginal assignments** of x by

$$y(i) := x(\{i\} \times O), \tag{5}$$
$$z(j) := x(N \times \{j\}). \tag{6}$$

There $y(i)$ corresponds to the payoff of player i in game (N, d) and $z(j)$ to the cost of player j in (O, c). Thus x is a feasible assignment implies $y \in core(N, d)$ and $z \in core(O, c)$, that is, the core of each marginal game must be non-empty.

Conversely, suppose the cores of the marginals games are non-empty. Pick any $y^* \in core(N, d)$ and $z^* \in core(O, c)$. Then one question is: does there exist a market assignment $x^* \in [0, 1]^E$ such that (y^*, z^*) are the marginal assignments of x^*?

Definition 3 Let $((N, d), (O, c))$ be a pair of marginal games (i.e., a whole game) with $d(N) = c(O)$. We say $(y, z) \in \mathbb{R}^N \times \mathbb{R}^O$ is a joint-core vector if

(i) y is in the core of game (N, d); and
(ii) z is in the core of game (O, c); and
(iii) (y, z) are the marginals of some assignment $x \in [0, 1]^E$.

The definition shows that the existence of a stable (i.e., core) allocation in each marginal game need not to imply a stable allocation for the whole game, as players in one game may not be able to implement their own allocation without the presence of the players in the other game. Therefore, the whole game requires a stronger condition to ensure stability (see Theorem 2).

The following result follows trivially from the above definitions and establishes an equivalence between quota implementability and non-emptyness of a joint core in the corresponding pair of marginal games.

[8] Here marginal games refer to marginal probability and should be distinguished from marginal product in the value theory.

Lemma 1 *Let (c, d) be a quota system with $d(N) = c(O)$. Then (c, d) is implementable if and only if the pair of marginal games $((N, d), (O, c))$ has a non-empty joint-core.*

3 Characterization

Theorem 1 completely characterizes the set of all implementable quotas with distributional and group influence constraints.

Theorem 1 *Suppose (c, d) is two-sided paramodular. (c, d) is implementable if and only if for all $A \subseteq N$ and $B \subseteq O$,*

$$d(A) - c(B) \leq |A||B^c|. \tag{7}$$

The characterization inequality (7) can be interpreted as a comparative balance between the quota floors of a set of agents and the quota ceilings of a set of goods. That is, the excess demand floor quota $d(A)$ over the supply ceiling quota $c(B)$ should not exceed the size of set A times the number of goods not in set B. This ensures that if there is an imbalance between demand and supply, enough non-quota-restricted goods are available which can be used to fulfill additional demand of agents in the set A.

We use a simple 3 agents-3 goods example to illustrate how Theorem 1 works when condition (7) holds, and by a minor change on the quotas, (7) does not hold, i.e., the implementability is upset.

Example 1 Let $N = \{i_1, i_2, i_3\}$ and $O = \{j_1, j_2, j_3\}$. Consider the following (agent) symmetric demand quotas and (good) symmetric supply quotas: for all $i, i' \in N$ and $j, j' \in O$,

$$\begin{aligned} &d(i) = 1, \quad d(i, i') = 3, \quad d(N) = 7, \\ &c(j) = 3, \quad c(j, j') = 5, \quad c(N) = 7. \end{aligned}$$

It is easy to verify that (c, d) is two-sided paramodular and satisfies all inequalities in (7). From Theorem 1, (c, d) is implementable, e.g., the allocation $x \in \mathbb{R}^9$ defined by $x(i_1, j_3) = 0$, $x(i_2, j_2) = 0$, and $x(i, j) = 1$ otherwise satisfies all the above quotas and hence is a feasible assignment. Consider modifying the supply quotas c by the following submodular quotas $\tilde{c}$:

$$\begin{aligned} &\tilde{c}(j_1) = 0, \quad \tilde{c}(j_2) = \tilde{c}(j_3) = 5, \\ &\tilde{c}(j_1, j_2) = \tilde{c}(j_1, j_3) = 5, \quad \tilde{c}(j_2, j_3) = \tilde{c}(N) = 7. \end{aligned}$$

We claim that $(\tilde{c}, d)$ violates condition (7). In particular, for testing set $(A, B) = (N, \{j_1\})$, the inequality

$$\tilde{c}(\{j_1\}) + |N||O \backslash \{j_1\}| \geq d(N)$$

is violated. Therefore, by Theorem 1 there exists no feasible assignment for $(\tilde{c}, d)$. Indeed, if there exists a feasible solution x, we must have $x(i, j_1) = 0$ for all $i \in N$ given $\tilde{c}(j_1) = 0$. But then $x(N \times O) \leq 6 < 7 = d(N)$. Contradiction.

Remark 2 Theorem 1 assumes that each agent demands at most one unit of each good. Our network flow characterization can be easily extended to the case of multiple-unit demands (see the proof of Theorem 1). Specifically, for each agent-good pair (i, j), let $u(i, j) \geq 0$ denote the upper quota. Then the upper bound $|A||B^c|$ in the characterization condition (7) can be modified as $u(A \times B^c)$.

Remark 3 A characterization of implementability with distributional constraints only appears to be difficult. Theorem 1 shows that a joint implementation of distributional constraints and side-constraints can lead to a simple characterization.

Combined with Lemma 1, Theorem 1 has game-theoretic interpretations. If (c, d) is two-sided paramodular, then (N, d) is a convex value game and (O, c) is a concave cost game. From Shapley (1971), the core of each game is non-empty (e.g., the Shapley value defined as the average of marginal values is in the core). Notice that this does not immediately ensure that the whole game has a non-empty joint-core. However, Theorem 1 characterizes the additional constraints on (c, d) for a joint-core to exist.

Theorem 2 *Let $((N, d), (O, c))$ be a pair of marginal games where (N, d) is a convex value game and (O, c) is a concave cost game. $((N, d), (O, c))$ has a non-empty joint-core if and only if (7) holds.*

3.1 Proof of Theorem 1

We use a network flow approach to derive the implementability condition in Theorem 1. We transform the implementability problem into an independent flow problem and invoke a maximum flow-minimum cut theorem to obtain a characterization of implementability. Below we outline the formulation of the implementation problem (c, d) as an independent flow problem (Lemma 2) and defer the remaining proof of Theorem 1 to the Appendix.

The flow network construction. We first review some basics about polymatroid and independent flows (see e.g., Fujishige 2005). Let U be a finite ground set and $f, g : 2^U \to \mathbb{R}$. We say the set $(U, f) := \{x \in \mathbb{R}_+^U : x(S) \leq f(S),\ \forall S \subseteq U\}$ is a polymatroid if f is submodular, and the set $(U, g) := \{x \in \mathbb{R}_+^U : x(S) \geq g(S),\ \forall S \subseteq U\}$ is a contrapolymatroid if g is supermodular. Consider a capacitated network

$$(G = (S^+ \cup S^-, A), \bar{c}, \underline{c}, (S^+, \rho^+), (S^-, \rho^-))$$

where G is a bipartite graph with a vertex set consisting of sources S^+ and sinks S^-, and A is the set of arcs from sources to sinks. We have a upper and lower capacity function $\bar{c}, \underline{c} : A \to \mathbb{R}_+$, a contrapolymatroid (S^+, ρ^+), and a polymatroid (S^-, ρ^-). For each subset of vertices $U \subset S^+ \cup S^-$, denote $\Delta^+(U)$ (and $\Delta^-(U)$) the set of arcs leaving (and entering) U. A function $\psi : A \to \mathbb{R}_+$ is called a *feasible independent*

flow if it satisfies

$$\sum_{a\in\Delta^+(S)} \psi(a) \geq \rho^+(S),\ \forall S \subseteq S^+, \tag{8}$$

$$\sum_{a\in\Delta^-(T)} \psi(a) \leq \rho^-(T),\ \forall T \subseteq S^-, \tag{9}$$

$$\underline{c}(a) \leq \psi(a) \leq \bar{c}(a),\ \forall a \in A. \tag{10}$$

We call (8)–(9) the polymatroidal boundary constraints on the source set S^+ and the sink set S^- and (10) the flow capacity constraint for each arc. The independent flow problem is to determine whether there exists a feasible flow satisfying (8)–(10).

We now formulate the implementation problem (c, d) as an independent flow problem. The problem defines an independent flow network $P = (G = (N \cup O, E), \bar{c}, \underline{c}, (N, d), (O, c))$. Here G is a complete bipartite graph with the source set N and the sink set O, and E consists of arcs from each $i \in N$ to $j \in O$. We define $\underline{c}(i, j) = 0$ and $\bar{c}(i, j) = 1$ for each arc $(i, j) \in E$. We define $\rho^+ = d$ and $\rho^- = c$. Since we assume $-d$ and c are submodular, (N, d) is a contrapolymatroid and (O, c) is a polymatroid. Lemma 2 below is central for the proof of Theorem 1.

Lemma 2 *(c, d) is implementable if and only if the independent flow problem $P = (G = (N \cup O, E), \bar{c}, \underline{c}, (N, d), (O, c))$ has a feasible flow.*

Remark 4 For one-sided auction markets, Che et al. (2013) construct a polymatroidal flow network for their implementation problem. While our problem cannot transform into a one-sided auction problem with paramodular constraint, it can be shown that every independent flow problem has an equivalent representation as a polymatroidal flow problem (see Fujishige 2005).

3.2 Example: Partition-generated demand quotas

An example of Theorem 1 is the case of demand floor quotas generated by some collection of target groups $\mathcal{S} \subseteq 2^N \backslash \{\emptyset\}$ that are partitional. We show that for this class of demand floor quotas, we can obtain a more tractable reduction of the set of implementable quotas.

Let $\mathcal{S} = \{S_i\}_{i\in\mathcal{I}}$ be a partition of N with $\mathcal{I}$ being the index set. A function $d : 2^N \to \mathbb{R}_+$ is defined as $\mathcal{S}$-generated, if $d(\emptyset) = 0$ and for every $A = \cup_{i\in\mathcal{I}} A_i$ with $A_i \subseteq S_i$ for each $i \in \mathcal{I}$, it holds $d(A) = \sum_{i\in\mathcal{I}} d(A_i)$. Intuitively, for any $\mathcal{S}$-generated demand floor quotas, the quotas can be supermodular within each target group but are additive across different target groups. It is easy to verify that if d is $\mathcal{S}$-generated by some partition $\mathcal{S}$, then d is supermodular. From Theorem 1, we obtain the following characterization for $\mathcal{S}$-generated demand floor quotas: (c, d) is implementable if and only if for all $(A_i)_{i\in\mathcal{I}}$ with each $A_i \subseteq S_i$ and all $B \subseteq O$,

$$\sum_{i\in\mathcal{I}} d(A_i) - c(B) \leq \sum_{i\in\mathcal{I}} |A_i||B^c|. \tag{11}$$

When both the demand and supply quotas are modular, e.g., $\mathcal{S} = \{\{1\}, \{2\}, \ldots, \{n\}\}$, Theorem 1 reduces to the following simpler characterization.

Corollary 1 *Suppose* (c, d) *is modular. Then* (c, d) *is implementable if and only if for all* $A \subseteq N$ *and* $B \subseteq O$,

$$\sum_{i \in A} d(i) - \sum_{j \in B} c(j) \leq |A||B^c|. \tag{12}$$

4 Supermodular floor polytope

Theorem 1 characterizes the set of all implementable quota systems. It also implies a characterization of supermodular and implementable demand floor quotas with supply ceilings being given. For any given ceiling quota system c, define

$$f(A) := \min_{B}[c(B) + |A||B^c|]. \tag{13}$$

We say f is symmetric if for each $A \subseteq N$, $f(A)$ only depends on the cardinality of A; and f is monotone if $S \subset T$ implies $f(S) \leq f(T)$. The following result is immediate.

Lemma 3 *The bound* f *given by* (13) *is submodular, monotone, and symmetric. Moreover,* $f(A) > 0$ *for all* $A \neq \emptyset$.

We define the cone of all supermodular demand floor quotas by

$$\mathcal{P} = \{d \in \mathbb{R}_+^{2^N} : d(S \cap T) + d(S \cup T) \geq d(S) + d(T),\ \forall\, S, T \subseteq N,\ d(\emptyset) = 0\}. \tag{14}$$

Then $\mathcal{P}$ is a pointed cone and hence finitely generated by its extreme rays. We define the set of all implementable demand floor quotas by

$$\mathcal{Q} = \{d \in \mathbb{R}_+^{2^N} : d(S) \leq\ f(S),\ \forall\, S \subseteq N,\ d(\emptyset) = 0\}. \tag{15}$$

Then $\mathcal{Q}$ is a hyperrectangle $0 \leq d \leq f$ (i.e., a box). We denote the set of all supermodular and implementable demand floor quotas by

$$\mathcal{F} = \mathcal{P} \cap \mathcal{Q}. \tag{16}$$

We call $\mathcal{F}$ a **supermodular floor polytope**.

In other words, for a policymaker that determines floor quotas, her feasible choice set is a supermodular floor polytope. It is worth noting that characterizing the extreme rays of supermodular cones appears to be a challenging task, as indicated by Shapley (1971). This implies that characterizing the extreme points of the supermodular floor polytope, as defined by inequalities (14) and (15), is non-trivial or even more

challenging. To develop some intuition on the structure of the extreme points of $\mathcal{F}$, consider the following two examples.

Example 2 Let $N = \{1, \ldots, n\}$ and let $S_1, S_2, S_3 \subset N$, with $S_i \cap S_j = \emptyset$, be three subgroups, with $|S_1| + |S_2| < |S_3|$, e.g., S_1 and S_2 are minority groups. Suppose only two minority groups have influence on quotas, both separately and jointly, while individuals or smaller coalitions within each minority group, as well as the majority group has no influences. The corresponding demand floors are defined by

$$d(A) = \begin{cases} d(S_1) & \text{if } S_1 \subseteq A, S_1 \cup S_2 \nsubseteq A, \\ d(S_2) & \text{if } S_2 \subseteq A, S_1 \cup S_2 \nsubseteq A, \\ d(S_1 \cup S_2) & \text{if } S_1 \cup S_2 \subseteq A, \\ 0 & \text{Otherwise.} \end{cases}$$

Then d is supermodular. Moreover, $(d(1), d(2), d(12)) := (d(S_1), d(S_2), d(S_1 \cup S_2))$ are the essential demand floor quotas. For a given f, let $(f(1), f(2), f(12)) := (f(S_1), f(S_2), f(S_1 \cup S_2))$. From the definition of d and Lemma 3, it is easy to check that $d(1) \leq f(1), d(2) \leq f(2), d(12) \leq f(12)$ implies $d \leq f$. So we can restrict attention to the problem in $\mathbb{R}^3$.

The box $\mathcal{Q} \subset \mathbb{R}^3$ has 8 vertices: $\{0, f(1)\} \times \{0, f(2)\} \times \{0, f(12)\}$. The supermodular cone $\mathcal{P} \subset \mathbb{R}^3$ is given by

$$-d(1) - d(2) + d(12) \geq 0, \tag{17}$$

$$d(1) \geq 0, \tag{18}$$

$$d(2) \geq 0. \tag{19}$$

We claim that the cone has 3 extreme rays (up to positive scaling):

$$(0, 0, 1), (1, 0, 1), (0, 1, 1).$$

To see this claim, notice that a nonzero $d \in \mathbb{R}^3$ is an extreme ray if and only if there are 2 linearly independent constraints in (17)–(19) binding at d. If (17) and (18) are binding, we get $d = (0, \alpha, \alpha)$, with $\alpha > 0$. If (17) and (19) are binding, we get $d = (\alpha, 0, \alpha)$. If (18) and (19) are binding, we get $d = (0, 0, \alpha)$.

We now determine the extreme points of $\mathcal{F}$. From the definition of extreme points, $d \in \mathbb{R}^3$ is an extreme point of $\mathcal{F}$ if there are 3 linearly independent constraints in (17)–(19) and $0 \leq d \leq f$ binding at d. Notice that each extreme ray of the cone intersects the faces of the box $0 \leq d \leq f$ at $d = 0$ and a unique nonzero point d'. Since there are 3 linearly independent constraints binding at d', it is an extreme point of $\mathcal{F}$. These points give the following 4 extreme points of $\mathcal{F}$:

$$(0, 0, 0), (0, 0, f(12)), (f(1), 0, f(1)), (0, f(2), f(2)).$$

There are two other extreme points of $\mathcal{F}$ that correspond to two vertices of the box:

$$(f(1), 0, f(12)), (0, f(2), f(12)).$$

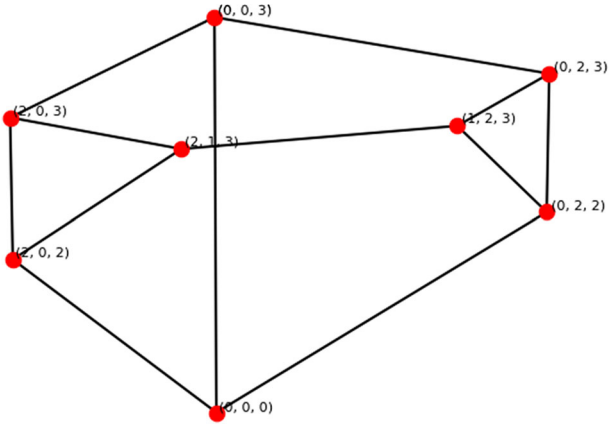

Fig. 1 Supermodular floor polytope

Finally, there are 2 extreme points that neither correspond to the extreme rays of the cone nor to the vertices of the box, but are generated by a combination of conic constraints and box constraints:

$$(f(1), f(12) - f(1), f(12)), (f(12) - f(2), f(2), f(12)).$$

Hence, there are 8 extreme points for $\mathcal{F}$ in total. For $f(1) = f(2) = 2$, $f(12) = 3$, the supermodular floor polytope is shown in Fig. 1.

The above example shows that some extreme points are generated by the intersection of the extreme rays of the supermodular cone and the box (type-I), while the remaining ones are either the extreme points of the box (type-II), or are generated by a combination of conic and box constraints (type-III). The following example shows that the number of type-I extreme points grows very quickly when the number of agents increases.

Example 3 Suppose $N = \{1, 2, 3, 4\}$. Then there are 15 possible non-empty subsets of N and $d \in \mathbb{R}_+^{15}$. It is well known that for $|N| = 4$, the supermodular cone has 37 extreme rays (Shapley 1971). These rays can be classified into 10 classes of permutably equivalent types. Using this result, we can find 10 (permutably equivalent) candidates of the extreme points generated by the intersection of the supermodular cone and the box. For illustration, we consider the first class of extreme rays (up to positive scaling) given by

S	1	2	3	4	12	13	14	23	24	34	123	124	134	234	1234
d^1	0	0	0	0	1	1	1	1	1	0	2	2	2	2	4

That is, the class contains 6 extreme rays that differ from d^1 only in the quotas of 2-agent coalitions, with one coalition's quota equal to 0 and other coalitions' quotas equal to 1.

To find the intersection of this extreme ray and the boundary of the box, below we show that for the box constraints $d(12) \leq f(12)$, $d(123) \leq f(123)$, and $d(1234) \leq f(1234)$, only the last inequality can be binding. We first claim that the first two cases are impossible for otherwise some other box constraint must be violated. Suppose $d(12) = f(12)$. Then d is on the ray d^1 implies

$$d(123) = 2d(12) = 2f(12).$$

On the other hand, the box constraints require

$$d(123) \leq f(123).$$

Since f is submodular, symmetric and strictly positive, e.g., $f(12) = f(13)$, and $f(1) > 0$, we have $2f(12) = f(12) + f(13) \geq f(1) + f(123) > f(123)$. Contradiction.

Similarly, we can rule out the case $d(123) = f(123)$. Finally, it can be verified that all box constraints can be satisfied when $d(1234) = f(1234)$. So we find one extreme point that is the intersection point of an extreme ray and the boundary of the box.

The above observation generalizes and we have the following partial characterization of the extreme points of $\mathcal{F}$. Before stating the theorem, we introduce some notations. Let $\mathbf{0}$ denote the zero vector in $\mathbb{R}^n$. Let $P \subset \mathbb{R}^n$ be a polytope. We say two extreme points of P are neighbors if the line segment connecting them is an edge (i.e., a one-dimensional face) of P. For each extreme point $v \in P$, let $N(v)$ denote the set of neighbors of v.

Theorem 3 (Extreme points of the supermodular floor polytope) *Let $\mathcal{F}$ be the supermodular floor polytope and let $\mathcal{P}$ be the supermodular cone in problem (P-b).*

(i) Every extreme ray of $\mathcal{P}$ contains exactly one non-zero extreme point of $\mathcal{F}$ that is a neighbor of $\mathbf{0}$.

(ii) Every non-zero extreme point of $\mathcal{F}$ that is a neighbor of $\mathbf{0}$ corresponds to an extreme ray of $\mathcal{P}$.

The theorem implies that in general there are two classes of extreme points of the supermodular floor polytope, namely, the non-zero extreme points that are neighbors of the zero vector and other non-zero extreme points that are not neighbors of the zero vector. For the first class, there is a one-to-one correspondence between the extreme points of the polytope and the extreme rays of the supermodular cone, while the second class corresponds to the extreme points of the box that are not eliminated by the supermodular constraints as well as some newly generated extreme points. For illustration, consider Example 2. The first three non-zero extreme points are the neighbors of the zero vector and hence they constitute the first class of the extreme points.

To prove Theorem 3, the following lemma will be useful for our analysis.

Lemma 4 (Ziegler 1995, p. 81) *Let $P \subset \mathbb{R}^n$ be a polytope, $v \in P$ be an extreme point, and let $N(v)$ be the set of its neighbors. Then the cone (based at v) spanned by the neighbors of v contains P: $P \subseteq v + cone\{u - v : u \in N(v)\}$.*

Proof of Theorem 3 (i) Let $\mathcal{P}^*$ be the set of the extreme rays of $\mathcal{P}$, which is unique up to positive scaling. Pick any $x \in \mathcal{P}^*$. Notice that for $\lambda > 0$ small, $\lambda x \in \mathcal{Q}$. So every extreme ray has a non-empty intersection with $\mathcal{Q}$. Define $\lambda^* = \max\{\lambda : \lambda x \in \mathcal{Q}\}$. Then λ^* is unique and $x^* := \lambda^* x \in \mathcal{F}$. We show that x^* is an extreme point of $\mathcal{F}$. Suppose not and there exist $x_1, x_2 \in \mathcal{F}$ and $\lambda \in (0, 1)$ such that $x^* = \lambda x_1 + (1-\lambda)x_2$. Then $x_1, x_2 \in \mathcal{P}$ and x^* is a nonnegative linear combination of x_1 and x_2. So x^* does not correspond to any extreme ray. Contradiction. Finally, $\lambda^* x$ is a neighbor of $\mathbf{0}$ since the line segment connecting $\mathbf{0}$ and $\lambda^* x$ is an edge of $\mathcal{F}$, i.e., for each interior point y on the segment, there are $n-1$ linearly independent constraints binding at y.

(ii) Let $\mathcal{N}$ be the cone spanned by the set $N(\mathbf{0})$ of non-zero extreme points of $\mathcal{F}$ that are neighbors of $\mathbf{0}$. We only need to show that $\mathcal{N} = \mathcal{P}$. First notice that $\mathcal{N} \subseteq \mathcal{P}$ since $N(\mathbf{0}) \subset \mathcal{F} \subset \mathcal{P}$. On the other hand, from Lemma 4, we have $\mathcal{F} \subseteq \mathcal{N}$. From (i), we have shown that each extreme ray $\lambda x \in \mathcal{P}$ corresponds to an extreme point $x^* = \lambda^* x \in \mathcal{F}$. Hence $x^* \in \mathcal{N}$. Since $\mathcal{N}$ is a cone, $\lambda x^* \in \mathcal{N}$ for all $\lambda > 0$. Hence $\mathcal{P} \subseteq \mathcal{N}$. We conclude that $\mathcal{N} = \mathcal{P}$. □

5 Target group-based quota design

In this section, we study a class of two-stage quota design games. At the first stage of a game, the planner chooses floor quotas with ceiling quotas being given. At the second stage, given the floor and ceiling quotas, some market mechanism runs and an assignment is determined. We assume the planner cannot determine mechanisms and assignments directly but can only influence the outcome through quotas. We discuss two examples: (1) an assignment problem with cardinal utilities and an efficient mechanism is used at the second stage; and (2) a many-to-one matching problem with ordinal preferences and a DA mechanism is used at the second stage.

5.1 Assignment market

We first extend our basic model in Sect. 2 to an assignment market, where the agents have monetary valuations over the goods.

Valuations. We assume that each agent $i \in N$ has a monetary valuation $v(i, j)$ for one unit of good $j \in O$, indicating the matching surplus between agent i and good j. For any feasible assignment $x \in \mathbb{R}^E$ and a group $S \subseteq N$, we can define the total surplus of group S as follows:

$$V_S(x) = \sum_{i \in S} \sum_{j \in O} v(i, j) x(i, j).$$

The planner's objective. We assume that the planner has a complete and transitive preference over assignments which can be represented by an objective function $W : \mathbb{R}^E \to \mathbb{R}$. There are certain target groups of agents, such as minority and majority groups. The planner values the surpluses of different target groups but with possibly different welfare weights. To formalize this model, let $\mathcal{S} \subset 2^N \setminus \{\emptyset\}$ be a collection

of target groups of agents, where each $S \in \mathcal{S}$ represents some agents with the same trait, e.g., age, sex, race, or medical vulnerability. Denote $l := |\mathcal{S}|$. Let $\lambda \in \mathbb{R}_+^{\mathcal{S}}$ denote social weights that assign to each target group $S \in \mathcal{S}$ a weight λ_S. One common scenario is when $\mathcal{S}$ consists of two complementary groups, such as a minority group $M \subset N$ and a majority group M^c. So a target group-based welfare objective can be parametrized by $(\mathcal{S}, \lambda)$.

Below are two examples of welfare objectives based on target groups.

1. The weighted utilitarian welfare. The planner maximizes a λ-weighted social surplus for the different target groups. The λ-weighted utilitarian objective is given by

$$U(\mathcal{S}, \lambda, x) = \sum_{S \in \mathcal{S}} \lambda_S V_S(x).$$

2. The weighted Nash's welfare. The different groups of agents bargain over the quotas and the surpluses, and the λ-weighted Nash product of the surpluses of these groups is maximized.[9] The λ-weighted Nash product is given by

$$N(\mathcal{S}, \lambda, x) = \prod_{S \in \mathcal{S}} [V_S(x)]^{\lambda_S}.$$

For welfare objectives with distributional concerns, Ashlagi and Shi (2016) discuss an allocation problem that maximizes a linear combination of utilitarian and max-min welfare objectives. Celebi and Flynn (2022) introduce three classes of welfare objectives in a school choice problem where agents' types determine their ordinal preferences and scores: (i) a λ-utilitarian objective that assigns different welfare weights to different types; (ii) an objective with a weighted score penalty function for different types; and (iii) an affirmative-action-concern objective with separable benefits for under-represented groups. Our specifications with target group-based objectives are similar to their first class of objectives.

Quota design game. Fix any ceiling quotas c and let $\mathcal{F} := \mathcal{F}(c)$ denote the set of implementable demand quotas given by Theorem 1 (and 3). Let $F(d) := F(c, d)$ denote the set of feasible assignments given any quotas (c, d). The quota design game has the following two stages:

In the second stage, given any d, we assume that the planner cannot directly influence the market allocation outcome, which assumes to be any socially efficient allocation $x^*(d)$ satisfying:

$$V_N(x^*(d)) = \max_{x \in F(d)} V_N(x). \qquad \text{(P-stage 2)}$$

In the first stage, the planner chooses a quota policy to maximize her welfare objective. Define $W^*(d) := W(x^*(d))$. The optimal quota problem is given by

$$\max_{d \in \mathcal{F}} W^*(d). \qquad \text{(P-stage 1)}$$

[9] Our model abbreviates the implementation of a welfare objective e.g., a Nash equilibrium in a Rubinstein bargaining game among different groups of agents.

Below, we provide some analysis on the structure of optimal quotas. Let $x, y \in \mathbb{R}^n$. We denote $x \geq y$ if $x_i \geq y_i$ for all $i = 1, \ldots, n$. Let $X \subset \mathbb{R}^n$. We define $f : X \to \mathbb{R}$ to be monotone (nonincreasing) if for any $x, y \in X$, the condition $x \geq y$ implies that $f(x) \leq f(y)$. We say y is minimal in X if there is no other vector x in X for which $x \leq y$. The following result provides a sufficient condition on the indirect welfare function such that the zero floor quotas $\mathbf{0} \in \mathcal{F}$ (e.g., laissez-faire) are an optimal solution.

Proposition 1 *If W^* is monotone in d on the feasible set $\mathcal{F}$, then $d^* = \mathbf{0}$ is an optimal solution to program* (P-stage 1).

When the indirect welfare function W^* is monotone, the zero vector is optimal when it is feasible. We note that the optimal solutions are not unique. There can be other solutions, including those that are not extreme points of $\mathcal{F}$ (See Example 4 below). In the case that the zero vector is not feasible, e.g., there are side-constraints $\mathcal{G}$ on floor quotas besides the distributional and group influence constraints $\mathcal{F}$ such that $\mathbf{0} \notin \mathcal{F} \cap \mathcal{G}$, it is easily seen that W^* remains monotone on this smaller feasible set, and hence a minimal floor vector in the set is optimal.

The above proposition implies the following sufficient condition for monotonicity.

Proposition 2 *If the planner's objective W is perfectly aligned with the second stage allocation rule (i.e., the efficient rule), then W^* is monotone.*

Proof of Proposition 2 Notice that in the second stage, decreasing each entry in d enlarges the set of feasible assignments, i.e., a feasible assignment remains feasible in the new problem with a reduced d. Hence, decreasing d will always weakly increase the value in the second stage, i.e., the second stage optimal value function is monotone. When the first stage objective W is equal to the second stage objective, W^* is also monotone. □

Generally, W^* may not be monotone. This non-monotonicity is due to the fact that when the second stage allocation rule is fixed as the efficient rule, the second stage always implements efficient allocations. When the planner wishes to implement a non-efficient allocation, setting zero floors cannot implement such allocations and strictly positive quotas are preferable. To illustrate non-monotone indirect welfare functions, let us consider a 3-agents 2-goods example.

Example 4 (Example 2 continued). Let $N = \{1, 2, 3\}$ and $O = \{a, b\}$, where $S_1 = \{1\}$, $S_2 = \{2\}$, and $S_3 = \{3\}$. Suppose $c(a) = c(b) = 2, c(ab) = 3$. Suppose it is common knowledge that agents 1 and 2 have a lower matching surplus than agent 3 for all goods (e.g., as when minority groups have lower average scores in school choice). Without loss, we assume $v(i, j) = v_i$ and $0 < v_1 < v_2 < v_3$. The second stage assignment problem is given by

$$\max_{0 \leq x \leq 1} \sum_{ij} v(i, j)x(i, j)$$
$$s.t.\ x(N \times \{j\}) \leq 2, \quad j = a, b$$

$$
\begin{aligned}
&x(N \times O) \leq 3\\
&x(\{i\} \times O) \geq d(i),\ i = 1, 2\\
&x(\{1, 2\} \times O) \geq d(12).
\end{aligned}
$$

From Theorem 1, we get $f(a) = f(b) = 2$, $f(ab) = 3$. Let $\mathcal{F}$ denote the set of supermodular and implementable demand floors. From Example 2, we know that the extreme points of $\mathcal{F}$ are given by $(0, 0, 0)$, $(2, 0, 2)$, $(0, 2, 2)$, $(0, 0, 3)$, $(2, 0, 3)$, $(0, 2, 3)$. We present all possible integral floors in $\mathcal{F}$, the corresponding efficient allocations and the λ-weighted utilitarian welfare in Table below:

$d = (d(1), d(2), d(12))$	$x^*(i, j) = 1$	$W(x^*)$
$(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 1)$	$2b, 3a, 3b$	$\lambda_2 v_2 + 2\lambda_3 v_3$
$(1, 0, 1)$	$1b, 3a, 3b$	$\lambda_1 v_1 + 2\lambda_3 v_3$
$(0, 0, 2)$, $(0, 1, 2)$, $(0, 2, 2)$	$2a, 2b, 3b$	$2\lambda_2 v_2 + \lambda_3 v_3$
$(1, 0, 2)$, $(1, 1, 2)$	$1b, 2b, 3b$	$\lambda_1 v_1 + \lambda_2 v_2 + \lambda_3 v_3$
$(2, 0, 2)$	$1a, 1b, 3b$	$2\lambda_1 v_1 + \lambda_3 v_3$
$(0, 0, 3)$, $(0, 1, 3)$, $(0, 2, 3)$	$1b, 2a, 2b$	$\lambda_1 v_1 + 2\lambda_2 v_2$
$(1, 0, 3)$, $(1, 1, 3)$, $(1, 2, 3)$		
$(2, 0, 3)$, $(2, 1, 3)$	$1a, 1b, 2b$	$2\lambda_1 v_1 + \lambda_2 v_2$

Notice that allocations implemented by $(1, 0, 1)$, $(1, 0, 2)$, and $(1, 1, 2)$, are not implementable by any extreme point of $\mathcal{F}$.

Consider the following three welfare objectives.

1. The utilitarian objective: $W = V_N$. That is, let $\mathcal{S} = \{N\}$ and $\lambda_N = 1$ in the λ-utilitarian welfare. Then W^* is monotone: decreasing each entry in d will enlarge the set of feasible assignments and increase the optimal value in the second stage (and hence the first stage). Hence, $(0, 0, 0)$ is an optimal solution in the first stage. In this case, the efficient allocation is given by $x^*(2b) = x^*(3a) = x^*(3b) = 1$. Moreover, the first stage problem has other solutions, such as $(0, 0, 1)$ and $(0, 1, 1)$, which are not extreme points of $\mathcal{F}$.

2. The λ-weighted utilitarian objective: $W = \lambda_1 V_1 + \lambda_2 V_2 + \lambda_3 V_3$, with each $\lambda_i > 0$. Consider outcome $x(1b) = x(3a) = x(3b) = 1$. The outcome is optimal if

$$\lambda_1 v_1 + 2\lambda_3 v_3 \geq W(x^*(d)),\ \text{ for all } d \neq (1, 0, 1),$$

which require $\lambda_3 v_3 \geq \lambda_1 v_1 \geq \lambda_2 v_2$. In this case, W^* is not monotone and $(1, 0, 1)$ is an optimal solution. Notice that this point is not an extreme point of $\mathcal{F}$. On the other hand, $x(1b) = x(2b) = x(3b) = 1$ is not implementable generically, since it requires $\lambda_3 v_3 = \lambda_1 v_1 = \lambda_2 v_2$.

3. The λ-weighted Nash objective: $W = V_1^{\lambda_1} V_2^{\lambda_2} V_3^{\lambda_3}$, with each $\lambda_i > 0$. Notice that for any allocation where only two agents receive goods, the utility of the remaining agent is zero and the Nash product is zero. On the other hand, the allocation $x(1b) = x(2b) = x(3b) = 1$ has a strictly positive Nash product $v_1^{\lambda_1} v_2^{\lambda_2} v_3^{\lambda_3}$. Hence, $(1, 0, 2)$ and $(1, 1, 2)$ dominate other integral quotas. We claim that an optimal solution can be fractional. Consider the following fractional quotas: $(\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$. The efficient allocation is given by $x(1b) = x(2b) = \frac{1}{4}, x(3a) = \frac{1}{2}, x(3b) = 1$. The Nash product is given by

$(\frac{1}{4}v_1)^{\lambda_1}(\frac{1}{4}v_2)^{\lambda_2}(\frac{3}{2}v_3)^{\lambda_3}$. It dominates $v_1^{\lambda_1}v_2^{\lambda_2}v_3^{\lambda_3}$ if $(\frac{1}{4})^{\lambda_1+\lambda_2}(\frac{3}{2})^{\lambda_3} \geq 1$. If we further assume $\lambda_1 + \lambda_2 + \lambda_3 = 1$, the condition reduces to $\lambda_3 \geq \ln 4/\ln 6$. In this case, all integral solutions are dominated and an optimal solution must be fractional.

Summarize (a)–(c), we conclude that depending on the type of welfare function, the indirect welfare function may be either monotone or non-monotone, the optimal quotas may be either extreme points or boundary points of the feasible set, and may be either integral or fractional.

5.2 Matching market

Below we consider a two-sided matching example where players have ordinal preferences. Consider a doctor-hospital matching problem with regional quotas, where H is a set of hospitals and D is a set of doctors. Each hospital $h \in H$ has a strict preference $\succ_h$ over the set of subsets of doctors. Each doctor $d \in D$ has a strict preference $\succ_d$ over the set of hospitals. Each hospital h has an upper capacity $c(h) \geq 0$ and a lower capacity $l(h) \geq 0$. We assume that $\succ_h$ is responsive (e.g., the ranking of a doctor is independent of other doctors). We further assume hospitals belong to different regions in $R = \{r_1, \ldots, r_n\}$ and each region $r \in R$ has a floor quota $l(r) \geq 0$ and a ceiling quota $c(r) \geq 0$ (e.g., regional caps).[10] The hospitals in the same region may influence the policy maker and change the floor quota.

A matching μ is a mapping that satisfies $\mu_d \in H \cup \{\emptyset\}$ for all $d \in D$, and $\mu_h \subseteq D$ for all $h \in H$, and for each pair (d, h), $\mu_d = h$ if and only if $d \in \mu_h$. A matching is **feasible** if $l(r) \leq |\cup_{h\in r} \mu_h| \leq c(r)$ for all $r \in R$, and $l(h) \leq |\mu_h| \leq c(h)$ for all $h \in H$. A matching is **individually rational** if (i) for each $d \in D$, $\mu_d \succeq_d \emptyset$, and (ii) for each $h \in H$, $d \succ_h \emptyset$ for all $d \in \mu_h$, and $|\mu_h| \leq c(h)$. That is, no player is matched with an unacceptable partner and each hospital's capacity is respected. We call a doctor-hospital pair (d, h) a **blocking pair** if $h \succ_d \mu_d$, and either (i) $|\mu(h)| < c(h)$ and $d \succ_h \emptyset$, or (ii) $d \succ_h d'$ for some $d' \in \mu_h$. We say a matching is **classically stable** if it is feasible, individually rational, and there is no blocking pair. We say a matching is **maximally stable** if it is feasible, individually rational, and minimizes the number of blocking pairs.

Gale and Shapley (1962) show that when there are only individual caps, a (classically) stable matching exists. However, when there are regional caps, a stable matching may not exist (see Kamada and Kojima (2015) for a discussion). Below we use an example to show that with floor quotas, the implementability of floor quotas is only necessary for classic stability. When the DA algorithm is used for the second stage, the solution may not be feasible. Moreover, there may exist no feasible and individually rational matching with no blocking pair. In particular, we show that when the quota policy is too extreme in the sense that only a certain group of hospitals is favored, no stable matching exists. In this case, we find the maximally stable matchings.

Example 5 Suppose there are three hospitals $H = \{h_1, h_2, h_3\}$ and three doctors $D = \{d_1, d_2, d_3\}$. We assume for each hospital i, $c(i) = 2$ and $l(i) = 0$. There

[10] The individual ceiling and floor quotas prevent overstaffing or understaffing at the hospital level and ensure that hospitals have sufficient staff to maintain quality care, while the regional ceiling and floor quotas ensure equitable distribution of doctors across different regions, especially in underserved or rural areas.

are two regions r_1 and r_2, with $h_1, h_2 \in r_1$ and $h_3 \in r_2$. The regions' floor quotas are $l(r_1), l(r_2) \geq 0$ and ceiling quotas are $c(r_1) = c(r_2) = 2$. The hospital and doctor rank order lists are given by

$$\begin{aligned} h_1 &: d_1, d_2, d_3 \qquad & d_1 &: h_2, h_1, h_3 \\ h_2 &: d_2, d_1, d_3 \qquad & d_2 &: h_1, h_3, h_2 \\ h_3 &: d_3, d_1, d_2 \qquad & d_3 &: h_3, h_1, h_2 \end{aligned}$$

The implementable demand floors are classified into three cases:

(1) $0 \leq l(r_1) \leq 2$ and $0 \leq l(r_2) \leq 1$. The doctor-proposing DA algorithm finds the doctor-optimal stable matching given by $\mu_0(d_1) = h_2$, $\mu_0(d_2) = h_1$ and $\mu_0(d_3) = h_3$,[11] which is feasible, individually rational, and with no blocking pair.

(2) $0 \leq l(r_1) \leq 1$ and $l(r_2) = 2$. The above DA matching is no longer feasible. We claim that there exists no stable matching. Note that in any stable matching, we must have d_3 is matched with h_3. Then consider the following four feasible and individually rational matchings:

$$\begin{aligned} &\mu_1(d_1) = h_1, \mu_1(d_2) = h_3, \mu_1(d_3) = h_3 \\ &\mu_2(d_1) = h_2, \mu_2(d_2) = h_3, \mu_2(d_3) = h_3 \\ &\mu_3(d_1) = h_3, \mu_3(d_2) = h_2, \mu_3(d_3) = h_3 \\ &\mu_4(d_1) = h_3, \mu_4(d_2) = h_2, \mu_4(d_3) = h_3. \end{aligned}$$

However, none of these matchings is stable, as one of the remaining doctors (d_1 or d_2) matched with h_3 and the unmatched hospital is a blocking pair. Among all feasible and individually rational matchings, the unique maximally stable matching is given by μ_2, in which (d_2, h_1) is the unique blocking pair.

(3) $l(r_1) = 3$ and $l(r_2) = 0$. Any feasible matching requires that d_3 is matched with either h_1 or h_2. Hence, none of these matchings is stable. We have the unique maximally stable matching is given by

$$\mu_5(d_1) = h_2, \mu_5(d_2) = h_1, \mu_5(d_3) = h_1.$$

When there are regional caps only, alternative notions of stability have been proposed. Kamada and Kojima (2015) introduce a stability weaker than classic stability in which certain types of blocking pairs are tolerated. They assume that there are some soft target capacities $(\bar{c}(h))_{h \in H}$ and regions prefer to respect target capacities as much as possible. Formally, a matching μ is stable if it is feasible, individually rational, and if (d, h) (with $h \in r$ for some r) is a blocking pair, then (i) $|\mu_r| = c(r)$, (ii) $d' \succ_h d$ for all $d' \in \mu_h$, and (iii) either $\mu_d \notin r$ or $|\mu'_h| - \bar{c}(h) > |\mu'_{\mu_d}| - \bar{c}(\mu_d)$, where μ' is modified from μ by moving d to h. That is, a blocking pair is legitimate if the movement of the doctor equalizes the excesses over the target capacities.[12]

[11] For notational convenience, we denote by $\mu(d)$ for μ_d.

[12] Kamada and Kojima (2018) generalize this stability concept to a model with general regional constraint structures and regional preferences. They show that hierarchical constraints is necessary and sufficient for the

To compare these alternative stability concepts, we study a simple example with one region, a regional cap and an individual floor constraint. Notice that the stability concept in Kamada and Kojima (2015) is defined in problems without floor constraints. Below we apply this notion with no modification to our problem.

Example 6 Suppose there are three hospitals $H = \{h_1, h_2, h_3\}$ and three doctors $D = \{d_1, d_2, d_3\}$. We assume for each hospital h, $c(h) = 2$, and $l(h_1) = l(h_2) = 0$, $l(h_3) = 1$. There is only one region $r_0 = H$, with $l(r_0) = 0$ and $c(r_0) = 2$. Let $\bar{c}(h) = 1$ for each $h \in H$ be the target capacities. The hospital and doctor rank order lists are given by

$$\begin{array}{ll} h_1 : d_1, d_2, d_3 & d_1 : h_1, h_2, h_3 \\ h_2 : d_2, d_1, d_3 & d_2 : h_1, h_2, h_3 \\ h_3 : d_3, d_1, d_2 & d_3 : h_2, h_3, h_1 \end{array}$$

The doctor-proposing DA algorithm finds the doctor-optimal stable matching:

$$\mu_0(d_1) = h_1, \mu_0(d_2) = h_1, \mu_0(d_3) = h_2.$$

The matching is not feasible as it violates the regional cap. It can be shown that there are two maximally stable matchings:

$$\begin{aligned} &\mu_1(d_1) = h_1, \mu_1(d_2) = \emptyset, \mu_1(d_3) = h_3, \\ &\mu_2(d_1) = \emptyset, \mu_2(d_2) = h_1, \mu_2(d_3) = h_3. \end{aligned}$$

Note that μ_1 and μ_2 are feasible and individually rational. For μ_1, there are four blocking pairs (d_2, h) for $h \in H$ and (d_3, h_2). For μ_2, there are four blocking pairs (d_1, h) for $h \in H$ and (d_3, h_2).

We claim that neither μ_1 nor μ_2 satisfies stability defined in Kamada and Kojima (2015). At μ_1, each (d_2, h) is a legitimate blocking pair: at μ_1 the regional cap is binding and the blocking fills a vacant position (their conditions (i) and (ii)). Also d_2 is assigned unmatched and blocking will violate the regional cap (the first case of condition (iii)). (d_3, h_2) is not a legitimate blocking pair, as $|\mu'_{h_2}| - \bar{c}(h_2) = |\mu'_{\mu_{h_3}}| - \bar{c}(\mu_{h_3}) = 0$, so (iii) does not hold. Hence μ_1 does not satisfy stability. At μ_2, (d_1, h_1) is not a legitimate blocking pair and hence μ_2 does not satisfy stability.[13]

existence of a stable and (doctor) strategy-proof mechanism. Notice that hierarchy implies submodularity but not converse. This implies that hierarchical regional caps are always implementable. On the other hand, this also implies that submodularity may not guarantee the existence of a stable and strategy-proof mechanism.

[13] μ_1 and μ_2 also do not satisfy stability defined in Kamada and Kojima (2018), in which a doctor-hospital pair is defined as illegitimate if the movement of the doctor in the pair does not lead to a Pareto superior distribution of doctors for the regions that control the two involving hospitals and have constraints binding. Stability requires that a blocking pair is tolerated if it is either infeasible or illegitimate. Consider μ_1 in our example. The pair (d_3, h_2) is not infeasible. Since r_0 controls $\mu_1(d_3) = h_3$ and h_2, and there is one layer of hierarchy, μ' with d_3 moving to h_2 cannot be Pareto superior to μ_1 for any subregions of r_0. Hence the pair (d_3, h_2) is not illegitimate. So (d_3, h_2) is not tolerated.

Hence, maximal stability is weaker than stability defined in Kamada and Kojima (2015) for this example.

5.3 Discussion on sequential design

1. Computational burden. Kumano and Kurino (2024) show that for their quota adjustment mechanism, both the first-step algorithm of running the DA mechanism for an arbitrary quota distribution and the second-step algorithm of finding a quota-adjustment stable improvement cycle are polynomial. Hence, the entire algorithm is polynomial.

To see the computational complexity of the two-stage game in Sect. 4, first consider the second-stage problem $W^*(d) = \max\{V_N(x)|x \in F(d)\}$, which is an LP problem. When quotas are two-sided paramodular, the second stage is an independent flow problem, which can be solved in polynomial time. The first stage involves maximizing a convex function $W^*(d)$ over the superomodular floor polytope $\mathcal{F}$ (the convexity follows from the fact that the optimal value of an LP is convex in its parameters). The problem is generally NP-hard because it involves finding the maximum over the extreme points of the polytope, which can be exponentially many.[14] However, our characterization of all extreme points of $\mathcal{F}$ (Theorem 3) reduces the problem to checking two classes of extreme points (the neighbors and the non-neighbors of the zero vector), each of which has a more tractable structure.

2. The policy objectives. It is worth noting that the two-stage game where quotas and allocations are determined sequentially reflects a situation where the society values different objectives (diversity vs. stability) in a lexicographical order, that is, the first-stage objective has a higher priority over the second-stage objective. We have shown that for the two-stage games, a quota policy can cause inefficient or unstable outcomes and render the second-stage market ceasing to implement its own objective. The conflict between different policy makers raises the question of how different policy makers can compromise and/or create a unified design for both the quotas and the market mechanism (see Afacan et al. (2024) and Kumano and Kurino (2024) for a discussion where both the quotas and matching outcomes are simultaneously determined).

6 Conclusion

This paper studies the implementation of quota policies in assignment markets with distributional and group influence constraints. We characterize the set of all implementable quota systems and provide a game-theoretic interpretation. We use this implementability condition to analyze optimal quota policies. Our results can be applied to assignment and matching markets where quotas are either fully or partially controlled by a planner or determined through a bargaining process among different groups of agents.

[14] Also notice that the valuation of $W^*(d)$ for each d is non-trivial for large instances (e.g., when d is continuous).

Several avenues exist for extending our model, including (1) incomplete information; (2) side constraints; and (3) indivisibility. In this paper, we focus on complete information and set aside incentive issues. While we consider only group influence constraints, our model can incorporate additional side constraints. Finally, our model assumes that goods are perfectly divisible, although in many assignment problems, resources are indivisible. To characterize the feasible assignments with indivisibility, we need to address the issue of decomposing a random assignment. It can be seen that our Theorem 1 covers the cases with integral quotas but quota design with integral quotas is more involved. We leave these problems for future research.

Funding This work was supported by the National Natural Science Foundation of China (NSFC72033004 and NSFC72003165); Zhejiang Provincial Philosophy and Social Science Planning Project (Project No. 20JDZD020).

Data availability No data was used for the research described in the article.

Declarations

Conflict of interest None.

Appendix: Missing proofs

Proof of Lemma 2 Suppose first that (c, d) is implementable, which means there exists a feasible assignment x. In the independent flow problem P, we define $\psi(i, j) = x(i, j)$ for all $(i, j) \in E$. Clearly, ψ is a feasible independent flow. Conversely, suppose that in problem P, there exists a feasible independent flow ψ. We can define $x(i, j) = \psi(i, j)$ for all $(i, j) \in E$. Then x is a feasible assignment for the original problem. This completes the proof of the lemma. □

To complete the proof of Theorem 1, we will utilize the following lemma (Fujishige 2005, p.171).

Lemma 5 *There exists a feasible independent flow satisfying* (8)–(10) *if and only if, for each* $U \subseteq S^+ \cup S^-$*, the following conditions hold:*

$$\rho^+(S^+ \cap U) - \rho^-(S^- \cap U) \leq \bar{c}(\Delta^+(U)) - \underline{c}(\Delta^-(U)), \tag{20}$$

and for $U = S^+ \cup S^-$*,*

$$0 \leq \bar{c}(\Delta^+(U)) - \underline{c}(\Delta^-(U)). \tag{21}$$

Proof of Lemma 5 We refer to Fujishige (2005) for a formal proof. Below we provide an intuitive argument. For a flow network $(S^+ \cup S^-, A), \bar{c}, \underline{c}, (S^+, \rho^+), (S^-, \rho^-))$, a subset of nodes $U \subseteq S^+ \cup S^-$ is called a cut. Note that for each cut U, $\bar{c}(\Delta^+(U))$ is the total capacity for the arcs leaving U, while $\underline{c}(\Delta^-(U))$ is the total demand for the arcs entering U. Also, $\rho^+(S^+ \cap U)$ is the total demand for the arcs

entering the nodes $S^+ \cap U$ that cut the source set, and $\rho^-(S^- \cap U)$ is the total capacity for the arcs leaving the nodes $S^- \cap U$ that cut the sink set.[15]

From a version of Gale's demand theorem, a feasible flow exists if and only if for each cut U, the total demand $\underline{c}(\Delta^-(U)) + \rho^+(S^+ \cap U)$ is no greater than the total capacity $\bar{c}(\Delta^+(U)) + \rho^-(S^- \cap U)$. That is, there exists a feasible flow if and only if (20) holds for all $U \subseteq S^+ \cup S^-$. □

Proof of Theorem 1 Combining Lemma 5 with Lemma 2, we conclude that (c, d) is implementable if and only if, for all $U \subseteq N \cup O$, (20) holds and, for $U = N \cup O$, (21) holds.

Now, for every $U = A \cup B$ for some $A \subseteq N$ and $B \subseteq O$, condition (20) can be equivalently expressed as follows: for all $A \subseteq N$ and $B \subseteq O$,

$$d(A) - c(B) \leq \bar{c}(A \times B^c) - \underline{c}(A^c \times B).$$

Finally, condition (21) is redundant in this context. This completes the proof of the theorem. □

Proof of Lemma 4 We refer to Ziegler (1995) for a formal proof. Below we provide a proof sketch for this result. Fix any extreme point $v \in P$. Let $H = \{x : a^\top x = b\}$ be a cutting hyperplane such that $a^\top v > b > a^\top v'$ for all other extreme points $v' \in P$. Let $Q = P \cap H$ and let $ext(Q)$ denote the extreme points of Q. It can be shown that each $v' \in N(v)$ is in a one-to-one correspondence with some $u \in ext(Q)$. On the other hand, each ray emanating from v to any other point $x \in P$ contains a point of Q. Hence $P \subseteq v + cone\{u - v : u \in Q\} \subseteq v + cone\{u - v : u \in ext(Q)\} = v + cone\{u - v : u \in N(v)\}$. □

References

Abdulkadiroğlu A, Sönmez T (2003) School choice: a mechanism design approach. Am Econ Rev 93(3):729–747

Afacan M, Dur U, van der Linden M (2024) Capacity design in school choice. Games Econom Behav 146:227–291

Ashlagi I, Shi P (2016) Optimal allocation without money: an engineering approach. Manage Sci 90(10–11):1789–1823

Athey S, Coey D, Levin J (2013) Set-asides and subsidies in auctions. Am Econ J Microecon 5(1):1–27

Austen-Smith D, Wright J (1994) Counteractive lobbying. Am J Polit Sci 38(1):25–44

Bobbio F, Carvalho M, Lodi A, Rios I, Torrico A (2024) Capacity planning in stable matching. Working paper

Bogomolnaia A, Moulin H (2001) A new solution to the random assignment problem. J Econ Theory 100(2):295–328

Border K (1991) Implementation of reduced form auctions: a geometric approach. Econometrica 59(4):1175–1187

Budish E, Che Y, Kojima F, Milgrom P (2013) Designing random allocation mechanisms: theory and applications. Am Econ Rev 103(2):585–623

Celebi O, Flynn J (2022) Priority design in centralized matching markets. Rev Econ Stud 89:1245–1277

Celebi O, Flynn J (2021) Priorities vs. quotas. Working paper, MIT. Available at SSRN 3562665

[15] Note that in a classic node-capacitated flow network, the demand and capacity are defined for the nodes in S^+ and S^-. Here we have the demand and capacity defined over the sets of arcs adjacent to these nodes.

Chan J, Eyster E (2003) Does banning affirmative action lower college student quality? Am Econ Rev 93(3):858–872
Chan J, Eyster E (2009) The distributional consequences of diversity-enhancing university admissions rules. J Law Econ Org 25(2):499–517
Che Y, Kim J, Mierendorff K (2013) Generalized reduced-form auctions: a network-flow approach. Econometrica 81(6):2487–2520
Echenique F, Yenmez B (2015) How to control controlled school choice. Am Econ Rev 105(8):2679–2694
Ehlers L, Hafalir I, Yenmez B, Yildirim M (2014) School choice with controlled choice constraints: hard bounds versus soft bounds. J Econ Theory 153:648–683
Fragiadakis D, Troyan P (2016) Improving matching under hard distributional constraints. Theor Econ 12(2):863–908
Fryer R, Loury G, Yuret T (2008) An economic analysis of color-blind affirmative action. J Law Econ Org 24(2):319–355
Fujishige S (2005) Submodular functions and optimization, 2nd edn. Elsevier, Oxford
Gale D, Shapley L (1962) College admissions and the stability of marriage. Am Math Mon 69:9–15
Goeree J, Kushnir A (2023) A geometric approach to mechanism design. J Polit Econ Microecon 1(2):321–347
Hafalir I, Yenmez B, Yildirim A (2013) Effective affirmative action in school choice. Theor Econ 8(2):325–363
Kamada Y, Kojima F (2015) Efficient matching under distributional constraints: theory and applications. Am Econ Rev 105(1):67–99
Kamada Y, Kojima F (2018) Stability and strategy-proofness for matching with constraints: a necessary and sufficient condition. Theor Econ 13:761–793
Kojima F (2012) School choice: impossibilities for affirmative action. Games Econom Behav 75(2):685–693
Kumano T, Kurino M (2024) Quota adjustment process. Working paper, Keio University
Lang X, Mishra D (2024) Symmetric reduced form voting. Theor Econ 16:605–634
Matthews S (1984) On the implementability of reduced form auctions. Econometrica 52(6):1519–1522
Nguyen T, Vohra R (2019) Stable matching with proportionality constraints. Oper Res 67(6):1503–1519
Pai M, Vohra R (2012) Auction design with fairness concerns: Subsidies vs. set-asides. Discussion Paper, University of Pennsylvania
Pathak P, Sönmez T, Ünver U, Yenmez B (2023) Fair allocation of vaccines, ventilators and antiviral treatments: leaving no ethical value behind in health care rationing. Manage Sci 70(6):3381–4165
Shapley L (1971) Cores of convex games. Int J Game Theory 1(1):11–26
Shapley L, Shubik M (1971) The assignment game i: the core. Int J Game Theory 1(1):111–130
Vohra R (2011) Mechanism design: a linear programming approach. Cambridge University Press, Cambridge
Ziegler G (1995) Lectures on Polytopes. Graduate Texts in Mathematics, vol 152. Springer, Berlin